~~[illegible]r Refere[illegible]~~

~~No[illegible]e taken~~

~~f[illegible]n thi[illegible]brary~~

APR 0 4 2005
WITHDRAWN

The New Encyclopædia Britannica

Volume 22

MACROPÆDIA

Knowledge in Depth

WITHDRAWN

FOUNDED 1768
15TH EDITION

Encyclopædia Britannica, Inc.
Jacob E. Safra, Chairman of the Board
Jorge Aguilar-Cauz, President

Chicago
London/New Delhi/Paris/Seoul
Sydney/Taipei/Tokyo

SAN BRUNO PUBLIC LIBRARY

First Edition	1768–1771
Second Edition	1777–1784
Third Edition	1788–1797
Supplement	1801
Fourth Edition	1801–1809
Fifth Edition	1815
Sixth Edition	1820–1823
Supplement	1815–1824
Seventh Edition	1830–1842
Eighth Edition	1852–1860
Ninth Edition	1875–1889
Tenth Edition	1902–1903

Eleventh Edition
© 1911
By Encyclopædia Britannica, Inc.

Twelfth Edition
© 1922
By Encyclopædia Britannica, Inc.

Thirteenth Edition
© 1926
By Encyclopædia Britannica, Inc.

Fourteenth Edition
© 1929, 1930, 1932, 1933, 1936, 1937, 1938, 1939, 1940, 1941, 1942, 1943, 1944, 1945, 1946, 1947, 1948, 1949, 1950, 1951, 1952, 1953, 1954, 1955, 1956, 1957, 1958, 1959, 1960, 1961, 1962, 1963, 1964, 1965, 1966, 1967, 1968, 1969, 1970, 1971, 1972, 1973
By Encyclopædia Britannica, Inc.

Fifteenth Edition
© 1974, 1975, 1976, 1977, 1978, 1979, 1980, 1981, 1982, 1983, 1984, 1985, 1986, 1987, 1988, 1989, 1990, 1991, 1992, 1993, 1994, 1995, 1997, 1998, 2002, 2003, 2005
By Encyclopædia Britannica, Inc.

© 2005
By Encyclopædia Britannica, Inc.

Britannica, Encyclopædia Britannica, Macropædia, Micropædia, Propædia, and the thistle logo are registered trademarks of Encyclopædia Britannica, Inc.

Copyright under International Copyright Union
All rights reserved.

No part of this work may be reproduced or utilized in any form or by any means, electronic or mechanical, including photocopying, recording, or by any information storage and retrieval system, without permission in writing from the publisher.

Printed in U.S.A.

Library of Congress Control Number: 2004110413
International Standard Book Number: 1-59339-236-2

Britannica may be accessed at http://www.britannica.com on the Internet.

CONTENTS

Muḥammad and the Religion of Islām

Islām is a major world religion belonging to the Semitic family; it was promulgated by the Prophet Muḥammad in Arabia in the 7th century AD. The Arabic term *islām*, literally "surrender," illuminates the fundamental religious idea of Islām—that the believer (called a Muslim, from the active particle of *islām*) accepts "surrender to the will of Allāh (Arabic: God)." Allāh is viewed as the sole God—creator, sustainer, and restorer of the world. The will of Allāh, to which man must submit, is made known through the sacred scriptures, the Qur'ān (Koran), which Allāh revealed to his messenger, Muḥammad. In Islām Muḥammad is considered the last of a series of prophets (including Adam, Noah, Abraham, Moses, Jesus, and others), and his message simultaneously consummates and completes the "revelations" attributed to earlier prophets.

Retaining its emphasis on an uncompromising monotheism and a strict adherence to certain essential religious practices, the religion taught by Muḥammad to a small group of followers spread rapidly through the Middle East to Africa, Europe, the Indian subcontinent, the Malay Peninsula, and China. Although many sectarian movements have arisen within Islām, all Muslims are bound by a common faith and a sense of belonging to a single community.

This article deals with the life of Muḥammad and his revelation, with the fundamental beliefs and practices of Islām, and with the connection of religion and society in the Islāmic world. The history of the various peoples who embraced Islām is covered in the article ISLĀMIC WORLD.

This article is divided into the following major sections:

THE FOUNDATIONS OF ISLĀM

Muḥammad: the Prophet and his message

LIFE AND WORKS

Muḥammad (in full, Abū al-Qāsim Muḥammad ibn 'Abd Allāh ibn 'Abd al-Muṭṭalib ibn Hāshim) was born in Mecca *c.* 570 after the death of his father, 'Abd Allāh. Muḥammad was at first under the care of his paternal grandfather, 'Abd al-Muṭṭalib. Because the climate of Mecca was considered to be unhealthful, he was given as an infant to a wet nurse from a nomadic tribe and spent some time in the desert. At six he lost his mother, Āminah of the clan of Zuhra, and at eight his grandfather. Though his grandfather had been head of the prestigious Hāshem (Hāshim) clan and was prominent in Mecca politics, he was probably not the leading man in Mecca, as some sources suggest. Muḥammad came under the care of the new head of the clan, his uncle Abū Ṭālib, and is reputed to have accompanied him on trading journeys to Syria. About 595, on such a journey, he was in charge of the merchandise of a rich woman, Khadījah of the clan of Asad, and so impressed her that she offered marriage. She is said to have been about 40, but she bore Muḥammad at least two sons, who died young, and four daughters, of whom the best known was Fāṭimah, the wife of Muḥammad's cousin 'Alī, who is regarded as Muḥammad's divinely ordained successor by the Shī'ah branch of Islām. Until Khadījah's death in 619, Muḥammad took no other wife. The marriage was a turning point in Muḥammad's life. By Arab custom, minors did not inherit, and therefore Muḥammad had no share in the property of his father or grandfather; but by his marriage he obtained sufficient capital to engage in mercantile activity on a scale commensurate with his abilities.

Prophetic call and early religious activity. Muḥammad appears to have been of a reflective turn of mind and is said to have adopted the habit of occasionally spending nights in a hill cave near Mecca. The poverty and misfortunes of his early life doubtless made him aware of tensions in Meccan society. Mecca, inhabited by the tribe of Quraysh (Koreish), to which the Hāshim clan belonged, was a mercantile centre formed around a sanctuary, the Ka'bah (Kaaba), which assured the safety of those who came to trade at the fairs. In the later 6th century there was extensive trade by camel caravan between the Yemen and the Mediterranean region (Gaza and Damascus), bringing goods from India and Ethiopia to the Mediterranean; and the great merchants of Mecca had obtained monopoly control of this trade. Mecca was thus prosperous, but most of the wealth was in a few hands. Tribal solidarity was breaking up; merchants pursued individual interests and disregarded their traditional duties to the unfortunate.

Muḥammad's vision

About 610, as he reflected on such matters, Muḥammad had a vision of a majestic being (later identified with the angel Gabriel) and heard a voice saying to him, "You are the Messenger of God." This marked the beginning of his career as messenger (or apostle) of God (*rasūl Allāh*), or Prophet (*nabī*). From this time, at frequent intervals until his death, he received "revelations"—that is, verbal messages that he believed came directly from God. Sometimes these were kept in memory by Muḥammad and his followers, and sometimes they were written down. About 650 they were collected and written in the Qur'ān (or Koran, the sacred scriptures of Islām), in the form that has endured. Muslims believe the Qur'ān is divine revelation, written in the words of God himself.

Muḥammad is said to have been perturbed after the vision and first revelation but to have been reassured by his wife, Khadījah. In his later experiences of receiving messages there was normally no vision. (Occasionally there were physical concomitants, such as perspiring on a cold day, and these gave rise to the suggestion, now agreed to be unwarranted, that he was an epileptic.) Sometimes he heard a noise like a bell but apparently never a voice. The essence of such an experience was that he found a verbal message in his heart—that is, in his conscious mind. With the help of Khadījah's Christian cousin Waraqah, he came to interpret these messages as in general identical with those sent by God through other prophets or messengers to Jews, Christians, and others and to believe that by the first great vision and by the receipt of the messages he was commissioned to communicate them to his fellow citizens and other Arabs. In addition to proclaiming the messages he received, Muḥammad must have offered explanations and expositions of them in his own words, as is evident in the large body of prophetic traditions that the community has preserved.

Soon he gathered some sympathetic friends who accepted his claim to be a prophet and joined him in common worship and prayers. These culminated in an act of prostration in which they touched the ground with their foreheads in acknowledgment of God's majesty—still a cardinal act in Islāmic worship. In about 613 Muḥammad began preaching publicly, and he and his followers spent their days together in the house of a young man named al-Arqam. It is probable that they sometimes worshipped together in the Ka'bah, a sanctuary of the Arab pagans.

Pagan religious milieu

The people of Mecca at the time nominally worshipped many gods, but few believed that man was dependent on supernatural powers. The merchants thought most things could be accomplished by wealth and by human planning. Some men regarded Allāh as a "high god" who stood above lesser deities. (Allāh, the Arabic word for God, is used by Christian Arabs as well as by Muslims.) The earliest passages of the Qur'ān revealed to Muḥammad emphasize the goodness and power of God as seen in nature and in the prosperity of the Meccans and call on the latter to be grateful and to worship "the Lord of the Ka'bah," who is thus identified with God. Gratitude is to be expressed in generosity with one's wealth and avoidance of niggardliness. As a sanction, men are warned that they will appear before God on the Last Day to be judged according to their deeds and assigned to heaven or hell. (The doctrines of the Qur'ān are examined later in this article.)

By proclaiming this message publicly, Muḥammad gained followers—39, it is said—before he entered the house of al-Arqam. The names of 70 followers are known prior to the appearance of opposition to the new religion, and there were probably more. Most were young men under 30 when they joined Muḥammad. They included sons and brothers of the richest men in Mecca, though they might be described as persons excluded from the most lucrative forms of commerce. A handful of Muḥammad's early followers are spoken of as "weak," which merely means that they were not of the tribe of Quraysh and so not effectively protected by any clan. The new religion was eventually called Islām—*i.e.,* "surrender [to the will of God]"—and its adherents were called Muslims—*i.e.,* "those who have surrendered"—though the Qur'ān speaks of them primarily as "the believers."

Opposition at Mecca. Although Muḥammad's preaching was basically religious, there was implicit in it a critique of the conduct and attitudes of the rich merchants of Mecca. Attempts were made to get him to soften his criticism by offering him a fuller share in trade and a marriage alliance with one of the wealthiest families, but he decisively rejected such offers. About 615 more active opposition appeared. Points in the message of the Qur'ān were questioned, such as the assertion that men would be resurrected before the Judgment. Commercial pressure was brought to bear on Muḥammad's supporters, and in some families there was mild persecution of junior members who followed him. It is sometimes suggested that the main reason for opposition was the merchants' fear that the new religion would destroy the recognition of the Ka'bah as a sanctuary, but this is unlikely. Certainly attacks on idols appeared in the Qur'ān, and Islām began to be characterized by the insistence that "there is no god but God" (Allāh), but no attack was made on the Ka'bah, and the idols mentioned had their chief shrines elsewhere.

Commercial pressure, boycott, and persecution

A leader of the opposition arose in the person of Abū Jahl, a contemporary of Muḥammad, who probably felt that the latter, despite his claim to be "only a warner" (of Judgment to come), was building a position of authority that might one day make him politically supreme in Mecca, because Arabs deeply respected the kind of wisdom or knowledge that Muḥammad clearly had. In about 616 Abū Jahl organized a boycott of the clan of Hāshim by the chief clans of Mecca, allegedly because the clan continued to protect Muḥammad and did not curb his preaching; but, since few of the clan were Muslims, other questions may have been involved. After three years the boycott lost momentum, perhaps because some of the participants found they were harming their own economic interests.

Both Muḥammad's wife, Khadījah, and his uncle Abū Ṭālib died in about 619, and another uncle, Abū Lahab, succeeded as head of the clan of Hāshim. He was closer to the richest merchants, and at their instigation he withdrew the protection of the clan from Muḥammad. This meant that Muḥammad could easily be attacked and therefore could no longer propagate his religion in Mecca. He left for the neighbouring town of aṭ-Ṭā'if, but the inhabitants were insufficiently prepared to receive his message, and he failed to find support. Having secured the protection of the head of another clan, he returned to Mecca. In 620 Muḥammad began negotiations with clans in Medina, leading to his emigration, or *hijrah,* there in 622.

It is difficult to assess the nature and extent of the persecution of the Muslims in Mecca. There was little physical violence, and that almost always within the family. Muḥammad suffered from minor annoyances, such as having filth deposited outside his door. The persecution is said to have led to the emigration of some of the Muslims to Ethiopia about 615, but they may have been seeking opportunities for trade or military support for Muḥammad. Some remained until 628, long after Muḥammad was established in Medina. Whatever the nature of the persecution, the Muslims were very bitter about it.

The emigration from Mecca to Medina. In the summer of 621, 12 men from Medina, visiting Mecca for the annual pilgrimage to the Ka'bah (still a pagan shrine), secretly professed themselves Muslims to Muḥammad and went back to make propaganda for him at Medina. At the pilgrimage in June 622 a representative party of 75 persons from Medina, including two women, not merely professed Islām but also took an oath to defend Muḥammad as they would their own kin. These are known as the two Pledges of al-ʿAqaba. Muḥammad now encouraged his faithful Meccan followers to make their way to Medina in small groups, and about 70 emigrated thus. The Meccans are said to have plotted to kill Muḥammad before he could leave. With his chief lieutenant he slipped away unperceived, used unfrequented paths, and reached Medina safely on September 24, 622. This is the celebrated *hijrah* (Latin Hegira), which may be rendered "emigration," though the basic meaning is the severing of kinship ties. It is the traditional starting point of Islāmic history. The Islāmic Era (AH or Anno Hegirae) begins on the first day of the Arabic year in which the *hijrah* took place—July 16, 622, in the Western calendar.

Medina was different from Mecca. It was an oasis in which date palms flourished and cereals could be grown. Agriculture had been developed by several Jewish clans, who had settled among the original Arabs, and they still had the best lands. Later Arab immigrants belonging to the tribes of al-Aws and al-Khazraj, however, were in a stronger position. The effective units among the Arabs were eight or more clans, but nearly all of these had become involved in serious feuds. Much blood had been shed in a battle in about 618, and peace was not fully restored. In inviting Muḥammad to Medina, many of the Arabs there probably hoped that he would act as an arbiter among the opposing parties. Their contact with the Jews may have prepared them for a messianic religious leader, who would deliver them from oppression and establish a kingdom in which justice prevailed.

The Constitution of Medina

A document has been preserved known as the Constitution of Medina. In its present form it is a combination of at least two earlier documents and is probably later than 627, but its main provisions are almost certainly those originally agreed upon between Muḥammad and the Muslims of Medina. In form the document creates a confederation on traditional Arab lines among nine groups—eight Arab clans and the emigrants from Mecca. Muḥammad is given no special position of authority, except that the preamble speaks of the agreement as made between "Muḥammad the prophet" and the Muslims now resident in Medina, and it is stated that serious disputes are to be referred to him. The Jewish groups had refused to acknowledge Muḥammad as prophet and in the document appear in a secondary character as attached to various Arab clans. For at least five years, Muḥammad had no direct authority over members of other clans, but, in the closing years of his life, the prestige of his military successes gave him almost autocratic power. The revelations he received at Medina frequently contained legal rules for the community of Muslims, but they dealt with political questions only rarely.

The first five years at Medina. The first 18 months at Medina were spent in settling down. Muḥammad was given a piece of land and had a house built, which eventually held apartments grouped around a central courtyard for each of his wives. The Muslims often joined Muḥammad at prayers in his home, which, after his death, became the mosque of Medina. The emigrants (*muhājirūn*, the men from Mecca) were at first guests of brother Muslims in Medina, but Muḥammad cannot have contemplated this situation continuing indefinitely. A few emigrants carried on trade in the local market run by a Jewish clan. Others, with the approval of Muḥammad, set out in normal Arab fashion on razzias (*ghazawāt*, "raids") in the hope of intercepting Meccan caravans passing near Medina on their way to Syria. Muḥammad himself led three such razzias in 623. They all failed, probably because traitors betrayed the Muslim movements to the enemy. At last, in January 624, a small band of men was sent eastward with sealed orders telling them to proceed to Nakhlah, near Mecca, and attack a caravan from Yemen. This they did successfully, and in doing so they violated pagan ideas of sanctity—thereby making the Meccans aware of the seriousness of the threat from Muḥammad.

About the same time there was a change in Muḥammad's general policy in important respects. One aspect was the "break with Jews"; instead of making concessions to the Jews in the hope of gaining recognition of his prophethood, he asserted the specifically Arabian character of the Islāmic religion. Hitherto the Muslims had faced Jerusalem in prayer, but a revelation now bade them face Mecca. Perhaps because of this change some Muslims of Medina were readier to support Muḥammad. In March 624 he was able to lead about 315 men on a razzia to attack a wealthy Meccan caravan returning from Syria. The caravan, led by Abū Sufyān, the head of the Umayyah clan, eluded the Muslims by devious routes and forced marches. Abū Jahl, the head of the Makhzūm clan, however, leading a supporting force of perhaps 800 men, wanted to teach Muḥammad a lesson and did not withdraw. On March 15, 624, near a place called Badr, the two forces found themselves in a situation, perhaps contrived by Muḥammad, from which neither could withdraw without disgrace. In the ensuing battle at least 45 Meccans were killed, including Abū Jahl and other leading men, and nearly 70 taken prisoner, while only 14 Muslims died. To Muḥammad this appeared to be a divine vindication of his prophethood, and he and all the Muslims were greatly elated.

The Battle of Badr and its consequences

In the flush of victory some persons in Medina who had satirized Muḥammad in verse were assassinated, perhaps with his connivance. He also made a minor disturbance an excuse for expelling the Jewish clan, which ran the market. This weakened his most serious opponent there, the "hypocrite" (*munāfiq*), or nominal Muslim, ʿAbd Allāh ibn Ubayy, who was allied with the local Jews. The remaining waverers among the Arabs probably became Muslims about this time. Thus the victory of Badr greatly strengthened Muḥammad. At the same time he was using marriage relationships to bring greater cohesion to the emigrants. Of his daughters, Fāṭimah was married to ʿAlī (later fourth caliph, or leader of the Islāmic community) and Umm Kulthūm to ʿUthmān (third caliph). He himself was already married to ʿĀʾishah, daughter of Abū Bakr (first caliph), and was now espoused also to Ḥafṣah, daughter of ʿUmar (second caliph), whose previous husband was one of the Muslims killed at Badr.

In the same year Muḥammad led larger Muslim forces on razzias against hostile nomadic tribes and had some success. Presumably, he realized that the Meccans were bound to try to avenge their defeat. Indeed, Abū Sufyān was energetically mobilizing Meccan power. On March 21, 625, he entered the oasis of Medina with 3,000 men. One of the features of Medina was a large number of small forts that were impregnable to Arab weapons and tactics. Muḥammad would have preferred the Muslims to retire to these; but those whose cereal crops were being laid waste persuaded him to go out to fight. By a night march with 1,000 men, he reached the hill of Uḥud on the further side of the Meccan camp. On the morning of March 23 the Meccan infantry attacked and was repulsed with considerable loss. As the Muslims pursued, the Meccan cavalry launched a flank attack after the archers guarding the Muslim left had abandoned their position. The Muslims were thrown into confusion. Some made for a fort and were cut down, but Muḥammad and the bulk of his force managed to gain the lower slopes of Uḥud, where they were safe from the cavalry. The Meccans, because of their losses, were unable to press home their advantages and without delay set out for home, while Muḥammad the next day made a show of pursuing. The battle produced neither a clear victor nor loser. In Badr and Uḥud together, the Meccans had killed about as many men as they had lost; but they had boasted that they would make the Muslims pay several times over, and they had not shown the degree of superiority appropriate to their leading position in Arabia. Muḥammad, though he had lost above 70 men, realized that this was a military reverse, not a defeat; but the confidence of the Muslims and perhaps his own had been struck a serious blow. If the victory of Badr was a sign of God's support, did Uḥud indicate that he had abandoned the Muslims? Muḥammad's faith soon overcame any momentary doubts, and he was gradually able to restore the confidence of his followers.

The Battle of Uḥud

For two years after Uḥud, both sides prepared for a decisive encounter. In the razzias Muḥammad led or sanctioned, he seems to have aimed at extending his own alliances and at preventing others from joining the Meccans. In at least two cases a small party of Muslims was tricked or ambushed, and most of their lives were lost. And another Jewish clan was expelled from Medina. At length, in April 627 Abū Sufyān led a great confederacy of 10,000 men against Medina. On this occasion Muḥammad had ordered the crops to be harvested and a trench to be dug to defend the main part of the oasis from the Meccan cavalry. For a fortnight the confederates besieged the Muslims. Attempts to cross the trench failed, and fodder for the horses was scarce, while Muḥammad's agents among the attackers fomented potential dissensions. Then, after a night of wind and rain the great army melted away. The Meccans had exerted their utmost might and had failed to dislodge Muḥammad, whose position was now greatly strengthened.

The siege of Medina

For more than two years now there had been opposition to Muḥammad in Medina, chiefly from ʿAbd Allāh ibn Ubayy and other so-called hypocrites (*munāfiqūn*) who had abandoned Muḥammad at Uḥud and who together had fostered disaffection. Shortly before the siege Muḥammad had a showdown with ʿAbd Allāh ibn Ubayy, who had joined in spreading slanders about Muḥammad's wife ʿĀʾishah. This confrontation revealed that ʿAbd Allāh had little support in Medina, and he became reconciled to Muḥammad. After the siege of Medina, Muḥammad attacked the Jewish clan of Qurayẓah, which had probably been intriguing against him. When they surrendered, the men were all executed and the women and children sold as slaves.

The winning of the Meccans. Muḥammad's farsightedness as a statesman is manifest in the policies he next adopted. He might have proceeded to crush the Meccans, and he indeed put economic pressure on them; but his main aim was to gain their willing adherence to Islām. He had already realized that, insofar as the Arabs became Muslims, it would be necessary to direct outward the energies expended on razzias against one another. There could be no question of Muslims raiding Muslims. It is noteworthy that his largest razzias, apart from the expeditions against the Meccans, were along the route to Syria followed by the Arab armies after his death (see ISLAMIC WORLD). He doubtless realized that the administrative skill of the Meccan merchants would be required for any expansion of his embryonic state.

In a dream Muḥammad saw himself performing the annual pilgrimage to Mecca, and in March 628 he set out to do so, driving sacrificial animals; but he was disappointed because no more than 1,600 men would accompany him. The Meccans were determined to prevent the Muslims from entering their town, so Muḥammad halted at al-Ḥudaybiyah, on the edge of the sacred territory of Mecca. After some critical days the Meccans made a treaty with Muḥammad. Hostilities were to cease, and the Muslims were to be allowed to make the pilgrimage to Mecca in 629. The orderly withdrawal showed how completely Muḥammad controlled his followers. Partly to reward this orderly conduct, Muḥammad two months later led the same force against the Jewish oasis of Khaybar, north of Medina. After a siege it submitted, but the Jews were allowed to remain on condition of sending half of the date harvest to Medina. Thus throughout 628 and 629 Muḥammad's power was growing, since success led more men to become Muslims, for the religious attraction of Islām was apparently supplemented by material motives.

The Treaty of al-Ḥudaybiyah

Meanwhile Mecca was in decline. Several leading men had emigrated to Medina and become Muslims. New leaders had taken over from Abū Sufyān but had accomplished little, although the treaty with Muḥammad had removed his pressure on their caravans. Shortly after the treaty, Muḥammad had married Umm Ḥabībah, a daughter of Abū Sufyān and a widow whose Muslim husband had died in Ethiopia. This led to an understanding with Abū Sufyān, who began to work for the peaceful surrender of Mecca. It was probably when he was in Mecca for the pilgrimage in March 629 that Muḥammad became reconciled with another uncle, al-ʿAbbās, and married his uncle's sister-in-law Maymūnah.

An attack by Meccan allies in about November 629 upon allies of Muḥammad led to the latter's denunciation of the treaty of al-Ḥudaybiyah. After secret preparations he marched on Mecca in January 630 with 10,000 men. Abū Sufyān and other leading Meccans went out to meet him and formally submitted, and Muḥammad promised a general amnesty. When he entered Mecca there was virtually no resistance. Two Muslims and 28 of the enemy were killed. A score of persons were specifically excluded from the amnesty, but some were later pardoned. Thus Muḥammad, who had left Mecca as a persecuted prophet, not merely entered it again in triumph but also gained the allegiance of most of the Meccans. Though he did not insist on their becoming Muslims, many soon did so.

The victorious entrance into Mecca

Muḥammad spent 15 to 20 days in Mecca settling various matters of administration. Idols were destroyed in the Kaʿbah and in some small shrines in the neighbourhood. To relieve the poorest among his followers, he demanded loans from some of the wealthy Meccans. When he marched east to meet a new threat, 2,000 Meccans went with him.

The closing years: the unification of Arabia. Ever since the *hijrah,* Muḥammad had been forming alliances with nomadic tribes. At first these were probably nonaggression pacts, but, when he was strong enough to offer protection, he made it a condition of alliance that the tribe should become Muslim. While in Mecca Muḥammad had word of a large concentration of hostile nomads, and he set out to confront them. A battle took place at Ḥunayn in which part of Muḥammad's army was put to flight, but he himself and some older Muslims stood firm. The enemy was finally routed, and their dependents and possessions were all captured. They were allowed to ransom wives and children, but their livestock was divided as booty.

Muḥammad was now militarily the strongest man in Arabia. Most tribes sent deputations to Medina seeking alliance. It is difficult to say how much of Arabia joined his alliance, for the inner politics of each tribe were complex, and in some cases the deputation might represent only a small section. Muḥammad benefitted from the defeat of the Persian Empire by the Byzantine (Christian) Empire (627–628), for, in the Yemen and in places on the Persian Gulf, minorities that had relied on Persian support against Byzantium now turned to Muḥammad instead.

March to the Syrian border. The greatest of all of Muḥammad's razzias occurred at the end of 630, when he took 30,000 men on a month's journey to the Syrian border. In this campaign he pioneered the invasion of Syria and made agreements that became models for treaty arrangements with captured peoples. Some of the tribes near Syria were Christian and adhered to the Byzantines; chiefly as a result of this, Muḥammad's earlier friendship for the Christians, notably those of Ethiopia, changed to hostility. Before his death, armed opposition to him appeared in one or two parts of Arabia, but the Islāmic state was strong enough to deal with this. Thus he left most of Arabia united and poised for expansion into Syria and Iraq.

Muḥammad personally led the pilgrimage to Mecca, in March 632, in a form according with Islāmic belief. Although he had been in poor health for some time, no arrangement had been made for the succession. Thus his death at Medina in June 632 provoked a major crisis among his followers. The dispute over the leadership of the Muslim community eventually resulted in the most important schism in the history of Islām. (This development is discussed later in this article; see below *Theology and sectarianism.*)

CHARACTER AND ACHIEVEMENTS

Although greatly maligned by medieval European scholars—whose opinions still retain some influence—Muḥammad came to be viewed more objectively in the 19th century. Some of the evidence against him, such as his connivance at assassinations and his approval of the execution of the men of a Jewish clan, are historical matters that cannot be denied.

By his contemporaries, however, Muḥammad was admired for his courage, resoluteness, and impartiality, and for a firmness that was tempered by generosity. He won men's hearts by his personal charm. He was gentle, especially with children. Though he was sometimes silent in thought, for the most part he was engaged in purposeful activity. He walked vigorously and spoke rapidly. He became for later Muslims an exemplar of virtuous character, and stories presented him as realizing the Islāmic ideal of human life.

Muḥammad's chief significance is as founder of a state and as the Prophet of Islām. In his lifetime he created a federation of Arab tribes, which, in less than 20 years after his death, defeated the Byzantine and Persian empires, occupied a vast territory from Libya to Persia, and then developed into the Arab, or Islāmic, Empire. He made the religion of Islām the basis of Arab unity. Islāmic doctrine maintains that God is the founder of the religion, not Muḥammad, but the latter played an obviously important part in fostering the nascent religion. His concern with ultimate questions, his mystical outlook, and his moral seriousness were important adjuncts to the preaching of the Qur'ānic message. (W.M.W./Ed.)

THE LEGACY OF MUḤAMMAD

From the beginning of Islām, Muḥammad inculcated a sense of brotherhood and a bond of faith among his followers, helping to develop among them a close relationship that was accentuated by their experiences of persecution as a nascent community in Mecca. The strong attachment to the tenets of the Qur'ānic revelation and conspicuous socioeconomic content of Islāmic religious practices cemented this bond of faith. In AD 622, when the Prophet went to Medina, his preaching was soon accepted, and the community-state of Islām emerged. During this early period, Islām acquired its characteristic ethos as a religion uniting in itself both the spiritual and temporal aspects of life and seeking to regulate not only the individual's relationship to God (through his conscience) but human relationships in a social setting as well. Thus, there is not only an Islāmic religious institution but also an Islāmic law, state, and other institutions governing society. Not until the 20th century were the religious (private) and the secular (public) distinguished by some Muslim thinkers and separated formally, as in Turkey.

This dual religious and social character of Islām, expressing itself in one way as a religious community commissioned by God to bring its own value system to the world through the *jihād* ("holy war" or "holy struggle"), explains the astonishing success of the early generations of Muslims. Within a century after the Prophet's death in AD 632, they had brought a large part of the globe—from Spain across Central Asia to India—under a new Arab Muslim empire.

Relationship to other religions

The period of Islāmic conquests and empire building marks the first phase of the expansion of Islām as a religion. Islām's essential egalitarianism within the community of the faithful and its official acceptance of the followers of other religions won rapid converts. Jews and Christians were assigned a special status as communities possessing scriptures and called the "people of the Book" (*ahl al-kitāb*) and, therefore, were allowed religious autonomy. They were, however, required to pay a per capita tax called *jizyah*, as opposed to pagans, who were required to either accept Islām or die. The same status of the "people of the Book" was later extended to Zoroastrians and Hindus, but many "people of the Book" joined Islām in order to escape the disability of the *jizyah*. A much more massive expansion of Islām after the 12th century was inaugurated by the Ṣūfīs (Muslim mystics), who were mainly responsible for the spread of Islām in India, Central Asia, Turkey, and sub-Saharan Africa (see below).

Besides the *jihād* and Ṣūfī missionary activity, another factor in the spread of Islām was the far-ranging influence of Muslim traders, who not only introduced Islām quite early to the Indian east coast and South India but who proved as well to be the main catalytic agents (besides the Ṣūfīs) in converting people to Islām in Indonesia, Malaya, and China. Islām was introduced to Indonesia in the 14th century, hardly having time to consolidate itself there politically before coming under Dutch colonial domination.

The vast variety of races and cultures embraced by Islām (estimated to total from 1.1 billion to 1.2 billion persons worldwide) has produced important internal differences. All segments of Muslim society, however, are bound by a common faith and a sense of belonging to a single community. With the loss of political power during the period of Western colonialism in the 19th and 20th centuries, the concept of the Islāmic community (*ummah*), instead of weakening, became stronger. The faith of Islām helped various Muslim peoples in their struggle to gain political freedom in the mid-20th century and the unity of Islām contributed to later political solidarity.

Sources of Islāmic doctrinal and social views

Islāmic doctrine, law, and thinking in general are based upon four sources, or fundamental principles (*uṣūl*): (1) the Qur'ān, (2) the *sunnah* ("traditions"), (3) *ijmāʿ* ("consensus"), and (4) *ijtihād* ("individual thought").

The Qur'ān (literally, Reading, or Recitation) is regarded as the Word, or Speech, of God delivered to Muḥammad by the angel Gabriel. Divided into 114 *sūrah*s (chapters) of unequal length, it is the fundamental source of Islāmic teaching. The *sūrah*s revealed at Mecca during the earliest part of Muḥammad's career are concerned mostly with ethical and spiritual teachings and the Day of Judgment. The *sūrah*s revealed at Medina at a later period in the career of the Prophet are concerned for the most part with social legislation and the politico-moral principles for constituting and ordering the community. *Sunnah* ("a well-trodden path") was used by pre-Islāmic Arabs to denote their tribal or common law; in Islām it came to mean the example of the Prophet; *i.e.*, his words and deeds as recorded in compilations known as Ḥadīth.

Ḥadīth, or collection of the Prophet's sayings

Ḥadīth (a Report, or collection of sayings attributed to the Prophet) provide the written documentation of the Prophet's word and deeds. Six of these collections, compiled in the 3rd century AH (9th century AD), came to be regarded as especially authoritative by the largest group in Islām, the Sunnites. Another large group, the Shīʿites, has its own Ḥadīth.

The doctrine of *ijmāʿ*, or consensus, was introduced in the 2nd century AH (8th century AD) in order to standardize legal theory and practice and to overcome individual and regional differences of opinion. Though conceived as a "consensus of scholars," in actual practice *ijmāʿ* was a more fundamental operative factor. From the 3rd century AH *ijmāʿ* has amounted to a principle of stable thought; points on which consensus was reached in practice were considered closed and further substantial questioning of them prohibited. Accepted interpretations of the Qur'ān and the actual content of the *sunnah* (*i.e.*, Ḥadīth and theology) all rest finally on the *ijmāʿ*.

Ijtihād ("to endeavour" or "to exert effort") was required to find the legal or doctrinal solution to a new problem. In the early period of Islām, because *ijtihād* took the form of individual opinion (*ra'y*), there was a wealth of conflicting opinions. In the 2nd century AH *ijtihād* was replaced by *qiyās* (reasoning by strict analogy), a formal procedure of deduction based on the texts of the Qur'ān and the Ḥadīth. The transformation of *ijmāʿ* into a conservative mechanism and the acceptance of a definitive body of Ḥadīth virtually closed the "gate of *ijtihād*" in Sunnī Islām while *ijtihād* continued in Shīʿism. Nevertheless, certain outstanding Muslim thinkers (*e.g.*, al-Ghazālī, died AD 1111) continued to claim the right of new *ijtihād* for themselves, and reformers of the 18th and 19th centuries have caused this principle to once more receive wider acceptance.

The Qur'ān and Ḥadīth are treated in the following sections. The significance of *ijmāʿ* and *ijtihād* are discussed below in the contexts of Islāmic theology, philosophy, and law. (F.R./Ed.)

Islāmic scripture: the Qur'ān

The Qur'ān (Arabic: Reading or Recitation; often spelled Koran), the holy book of Islām, is regarded by believers as the true word of God as revealed to the Prophet Muḥammad. In its written form it is accepted as the earthly reproduction of an uncreated and eternal heavenly original, according to the general view referred to in the Qur'ān itself as "the well-preserved tablet" (*al-lawḥ al-maḥfūẓ*; Qur'ān 75:22). The word *qur'ān* is derived from the verb *qara'a* "to read," "to recite," but there is probably also some connection with Syriac *qeryānā*, "reading," used for the scriptural lessons in the Syrian Church. In the Qur'ān itself the word is not used with reference to the book as a whole but only as a term for separate revelations or for the divine revelation in general. The Qur'ān is held in high esteem as the ultimate authority in all matters legal and religious and is generally regarded as infallible in all respects. Its Arabic language is thought to be unsurpassed in purity and beauty and to represent the highest ideal of style. To imitate the style of the Qur'ān is a sacrilege.

FORM

Division into chapters and verses

In length the Qur'ān is approximately comparable with the New Testament. For purposes of recitation during the holy month of Ramaḍān it is divided into 30 "portions" (*juz'*, plural *ajzā'*), one for each day of the month. Its main division, however, is into 114 chapters, called *sūrah*s, of very unequal length. With the exception of the first *sūrah*, the so-called *fātiḥah* ("opening" of the book), which is a short prayer, the *sūrah*s are arranged roughly according to length, *sūrah* 2 being the longest and the last two or three the shortest. Because the longest *sūrah*s generally derive from the latter part of Muḥammad's activity, the consequence of this arrangement is that the oldest *sūrah*s are generally to be found toward the end of the book and the youngest generally appear at its beginning.

In the accepted version of the Qur'ān now in use, each *sūrah* has a heading containing the following elements: (1) a title, which is usually derived from some conspicuous word in the *sūrah*, such as "The Cow," "The Bee," "The Poets," but is usually not an indication of the contents of the whole chapter; (2) the *basmalah; i.e.*, the formula-prayer "In the name of God, the Merciful, the Compassionate"; (3) an indication of whether the *sūrah* was revealed at Mecca or at Medina and of the number of its verses; and finally (4) in some cases one or more *fawātiḥ*, or detached letters (*e.g.*, *ṭā' sīn, ṭā' sīn mīm*), or *alif lām mīm*, the meaning of which has not been satisfactorily explained, though it is thought that they might stand for abbreviated words, indicate certain collections of *sūrah*s, or have an esoteric significance.

The verses in the Qur'ān are called *āyah* (plural *āyāt*, literally "signs") and vary considerably in length. The shortest verses generally occur in the earliest *sūrah*s, in which the style of Muḥammad's revelation comes very close to the rhymed prose (*saj'*) used by the *kāhin*s, or soothsayers, of his time. As the verses get progressively longer and more circumstantial, the rhymes come farther and farther apart. There is also a change of linguistic style: the earlier *sūrah*s are characterized by short sentences, vivid expressions, and poetic force; and the later ones become more and more detailed, complicated and, at times, rather prosaic in outlook and language. As a result, it is sometimes difficult to decide whether or not a rhyme is intended to indicate the end of a verse; and consequently, there are variations in the numbering of verses (*e.g.*, between the European editions long used by Western scholars and the official Egyptian edition that has now replaced them in most scholarly works).

The Qur'ān generally appears as the speech of God, who mostly speaks in the first person plural ("we"). When the prophet Muḥammad is speaking to his compatriots, his words are introduced by the command, "Say," thus emphasizing that he is speaking on divine injunction only. At times the form is also dramatic, bringing in objections by Muḥammad's opponents and answering them by counter-arguments. Narrative passages are mostly brief. Stories of prophets and biblical persons are often alluded to as though they are known to the audience. The stress is not on the narrative but on its didactic uses.

On closer analysis very few of the *sūrah*s turn out to be uniform in style or content. The longest text dealing with one subject is *sūrah* 12, which tells the story of Joseph, differing from the biblical account in a great many details, most of which seem to outside historians to have been drawn from Jewish sources. Otherwise the longer *sūrah*s are composed of several brief sections dealing with a variety of topics. Thus the Qur'ān does not give the appearance of a planned, organized, or systematic treatise, an impression that is further heightened by the fact that certain favourite phrases such as "but God is forgiving, compassionate," "God is knowing, wise," "most of them know nothing" often have little or no apparent connection with the immediate context. In fact, some skeptics claim that these additions served only to produce a needed rhyme.

Heterogeneous style

It is often emphasized that Muḥammad brought to his people "an Arabic Qur'ān"; *i.e.*, a book or set of recitations in the Arabs' own language comparable to those of Judaism and Christianity. Also the vocabulary of the Qur'ān is overwhelmingly of Arabic origin, but there are, nevertheless, borrowed words, mostly from Hebrew and Syriac, bearing witness to Muḥammad's debt to Judaism and Christianity. These loan words are primarily technical terms such as *injīl*, "gospel" (Greek *evangelion*); *taurāt*, "the law, or Torah" of Judaism; *Iblīs*, "the Devil" (Greek *diabolos*); or translations or adaptations of theological terms such as *āmana*, "to believe" (Hebrew or Aramaic); *ṣalāt*, "prayer" (probably Syriac). Such explanations are usually regarded with suspicion by Muslims, since orthodox doctrine holds that the language of the Qur'ān is the purest Arabic. (H.R./Ed.)

DOCTRINES OF THE QUR'ĀN

God. The doctrine about God in the Qur'ān is rigorously monotheistic: God is one and unique; he has no partner and no equal. Trinitarianism, the Christian belief that God is three persons in one substance, is vigorously repudiated. Muslims believe that there are no intermediaries between God and the creation that he brought into being by his sheer command: "Be." Although his presence is believed to be everywhere, he does not inhere in anything. He is the sole Creator and sustainer of the universe, wherein every creature bears witness to his unity and lordship. But he is also just and merciful: his justice ensures order in his creation, in which nothing is believed to be out of place, and his mercy is unbounded and encompasses everything. His creating and ordering the universe is viewed as the act of prime mercy for which all things sing his glories. The God of the Qur'ān, described as majestic and sovereign, is also a personal God; he is viewed as being nearer to man than man's jugular vein, and, whenever a person in need or distress calls him, he responds. Above all, he is the God of guidance and shows everything, particularly man, the right way, "the straight path."

The God of the Qur'ān

This picture of God—wherein the attributes of power, justice, and mercy interpenetrate—is related to the Judeo-Christian tradition, whence it is derived with certain modifications, and also to the concepts of pagan Arabia, to which it provided an effective answer. The pagan Arabs believed in a blind and inexorable fate over which man had no control. For this powerful but insensible fate the Qur'ān substituted a powerful but provident and merciful God. The Qur'ān carried through its uncompromising monotheism by rejecting all forms of idolatry and eliminating all gods and divinities that the Arabs worshipped in their sanctuaries (ḥarams), the most prominent of which was Ka'bah sanctuary in Mecca itself.

The universe. In order to prove the unity of God, the Qur'ān lays frequent stress on the design and order in the universe. There are no gaps or dislocations in nature. Order is explained by the fact that every created thing is endowed with a definite and defined nature whereby it falls into a pattern. This nature, though it allows every created thing to function in a whole, sets limits; and this idea of the limitedness of everything is one of the most fixed points in both the cosmology and theology of the Qur'ān. The universe is viewed, therefore, as autonomous, in the sense that everything has its own inherent laws of behaviour, but not as autocratic, because the patterns of behaviour have been endowed by God and are strictly limited. "Everything has been created by us according to a measure." Though every creature is thus limited and "measured out" and hence depends upon God, God alone, who reigns unchallenged in the heavens and the earth, is unlimited, independent, and self-sufficient.

Man. According to the Qur'ān, God created two apparently parallel species of creatures, man and *jinn*, the one from clay and the other from fire. About the *jinn*, however, the Qur'ān says little, although it is implied that the *jinn* are endowed with reason and responsibility but are more prone to evil than man. It is with man that the Qur'ān, which describes itself as a guide for the human race, is centrally concerned. The Judeo-Christian story of the Fall of Adam (the first man) is accepted, but the Qur'ān states that God forgave Adam his act of disobedience, which is not viewed in the Qur'ān (in contradistinction to its understanding in the Christian doctrine) as original sin.

In the story of man's creation, angels, who protested to God against the creation of man, who "would sow mischief on earth," lost in a competition of knowledge against Adam. The Qur'ān, therefore, declares man to be the noblest of all creation, the created being who bore the trust (of responsibility) that the rest of the creation refused to accept. The Qur'ān thus reiterates that all nature has been made subservient to man: nothing in all creation has been made without a purpose, and man himself has not been created "in sport," his purpose being service and obedience to God's will.

Despite this lofty station, however, the Qur'ān describes human nature as frail and faltering. Whereas everything in the universe has a limited nature, and every creature recognizes its limitation and insufficiency, man is viewed as rebellious and full of pride, arrogating to himself the attributes of self-sufficiency. Pride, thus, is viewed as the cardinal sin of man, because by not recognizing in himself his essential creaturely limitations he becomes guilty of ascribing to himself partnership with God (*shirk:* associating a creature with the Creator) and of violating the unity of God. True faith (*īmān*), thus, consists of belief in the immaculate Divine Unity and Islām in one's submission to the Divine Will.

Pride, the cardinal sin

Satan, sin, and repentance. In order to communicate the truth of the Divine Unity, God has sent messengers or prophets to men, whose weakness of nature makes them ever prone to forget or even willfully reject the Divine Unity under the promptings of Satan. According to the Qur'ānic teaching, the being who became Satan (Shayṭān or Iblīs) had previously occupied a high station but fell from divine grace by his act of disobedience in refusing to honour Adam when he, along with other angels, was ordered to do so. Since then, his work has been to beguile man into error and sin. Satan is, therefore, the contemporary of man, and Satan's own act of disobedience is construed by the Qur'ān as the sin of pride. Satan's machinations will cease only on the Last Day.

Judging from the accounts of the Qur'ān, the record of man's accepting the prophets' messages has been rather dismal. The whole universe is replete with signs of God; the human soul itself is viewed as a witness of the unity and grace of God. The messengers of God have, throughout history, been calling man back to God. Yet very few men have accepted the truth; most of them have rejected it and become disbelievers (*kāfir*, plural *kuffār:* literally "ungrateful"—*i.e.*, to God), and when man becomes so obdurate, his heart is sealed by God. Nevertheless, it is always possible for a sinner to repent (*tawbah*) and redeem himself by a genuine conversion to the truth. There is no point of no return, and God is always willing and ready to pardon. Genuine repentance has the effect of removing all sins and restoring a person to the state of sinlessness with which he started his life.

Prophecy. Prophets are men specially elected by God to be his messengers. Prophethood is indivisible, and the Qur'ān requires recognition of all prophets as such without discrimination. Yet they are not all equal, some of them being particularly outstanding in qualities of steadfastness and patience under trial. Abraham, Noah, Moses, and Jesus were such great prophets. As vindication of the truth of their mission, God often vests them with miracles: Abraham was saved from fire, Noah from the deluge, and Moses from the Pharaoh. Not only was Jesus born from the Virgin Mary, but God also saved him from crucifixion at the hands of the Jews. The conviction that God's messengers are ultimately vindicated and saved is an integral part of the Qur'ānic doctrine.

All prophets are human and never part of divinity: they are simply recipients of revelation from God. God never speaks directly to a human: he either sends an angel messenger to him or makes him hear a voice or inspires him. Muḥammad is accepted as the last prophet in this series and its greatest member, for in him all the messages of earlier prophets were consummated. He had no miracles except the Qur'ān, the like of which no human can produce. (Soon after the Prophet's death, however, a plethora of miracles was attributed to him by Muslims.) The angel Gabriel brought the Qur'ān down to the Prophet's "heart." Gabriel is represented by the Qur'ān as a spirit, but the Prophet could sometimes see and hear him. According to early traditions, the Prophet's revelations occurred in a state of trance when his normal consciousness was in abeyance. This state was accompanied by heavy sweating. The Qur'ān itself makes it clear that the revelations brought with them a sense of extraordinary weight: "If we were to send this Qur'ān down on a mountain, you would see it split asunder out of fear of God."

Muḥammad as the last prophet

This phenomenon at the same time was accompanied by an unshakable conviction that the message was from God, and the Qur'ān describes itself as the transcript of a heavenly "Mother Book" written on a "Preserved Tablet." The conviction was of such an intensity that the Qur'ān categorically denies that it is from any earthly source, for in that case it would be liable to "manifold doubts and oscillations."

Eschatology. In Islāmic doctrine, on the Last Day, when the world will come to an end, the dead will be resurrected and a judgment will be pronounced on every person in accordance with his deeds. Although the Qur'ān in the main speaks of a personal judgment, there are several verses that speak of the resurrection of distinct communities that will be judged according to "their own book." In conformity with this, the Qur'ān also speaks in several passages of the "death of communities," each one of which has a definite term of life. The actual evaluation, however, will be for every individual, whatever the terms of reference of his performance. In order to prove that the resurrection will occur, the Qur'ān uses a moral and a physical argument. Because not all requital is meted out in this life, a final judgment is necessary to bring it to completion. Physically, God, who is all-powerful, has the ability to destroy and bring back to life all creatures, who are limited and are, therefore, subject to God's limitless power.

According to strict Qur'ānic doctrine, there is no intercession, although God himself, in his mercy, may forgive certain sinners. Those condemned will burn in hellfire, and those who are saved will enjoy the abiding pleasures of paradise. Hell and heaven are both spiritual and physical. Besides suffering in physical fire, the damned will also experience fire "in their hearts"; similarly, the blessed, besides physical enjoyment, will experience the greatest happiness of divine pleasure. Quite early, however, Islāmic tradition developed the notion of intercession, probably in answer to the Christian doctrine of redemption.

Social service. Because the purpose of the existence of man, as of every other creature, is submission to the Divine Will, God's role in relation to man is that of the commander. Whereas the rest of nature obeys God automatically, man alone possesses the choice to obey or disobey. With the deep-seated belief in Satan's existence, man's fundamental role becomes one of moral struggle, which constitutes the essence of human endeavour. Recognition of the unity of God does not simply rest in the intellect but entails consequences in terms of the moral struggle, which consists primarily in freeing oneself of narrowness of mind and smallness of heart. One must go out of oneself and expend one's best possessions for the sake of others.

God in the role of commander

The doctrine of social service, in terms of alleviating suffering and helping the needy, constitutes an integral part of the Islāmic teaching. Praying to God and other religious acts are deemed to be a pure facade in the absence of active welfare service to the needy. In regard to this matter, the Qur'ānic criticisms of human nature become very sharp: "Man is by nature timid; when evil befalls him, he panics, but when good things come to him he prevents them from reaching others." It is Satan who whispers into man's ears that by spending for others he will become poor. God, on the contrary, promises prosperity in exchange for such expenditure, which constitutes a credit with God and grows much more than the money people invest in usury. Hoarding of wealth without recognizing the rights of the poor is threatened with the direst punishment in the hereafter and is declared to be one of the main causes of the decay of societies in this world. The practice of usury is forbidden.

The concept of a community of the faithful

With this socioeconomic doctrine cementing the bond of faith, the idea of a closely knit community of the faithful who are declared to be "brothers unto each other" emerges. Muslims are described as "the middle community bearing witness on mankind," "the best community produced for mankind," whose function it is "to enjoin good and forbid evil" (Qur'ān). Cooperation and "good advice" within the community are emphasized, and a person who deliberately tries to harm the interests of the community is to be given exemplary punishment. Opponents from within the community are to be fought and reduced with armed force, if issues cannot be settled by persuasion and arbitration.

Because the mission of the community is to "enjoin good and forbid evil" so that "there is no mischief and corruption" on earth, the doctrine of *jihād,* in view of the constitution of the community as the power base, is the logical outcome. For the early community it was a basic religious concept. *Jihād*, or holy war, means an active struggle using armed force whenever necessary. The object of *jihād* is not the conversion of individuals to Islām but rather the gaining of political control over the collective affairs of societies to run them in accordance with the principles of Islām. Individual conversions occur as a by-product of this process when the power structure passes into the hands of the Muslim community. In fact, according to strict Muslim doctrine, conversions "by force" are forbidden, because after the revelation of the Qur'ān "good and evil have become distinct," so that one may follow whichever one may prefer (Qur'ān), and it is also strictly prohibited to wage wars for the sake of acquiring worldly glory, power, and rule. With the establishment of the Muslim empire, however, the doctrine of the *jihād* was modified by the leaders of the community. Their main concern had become the consolidation of the empire and its administration, and thus they interpreted the teaching in a defensive rather than in an expansive sense. The Khārijite sect (see below *Theology and sectarianism*) which held that "decision belongs to God alone," insisted on continuous and relentless *jihād,* but its followers were virtually destroyed during the internecine wars in the 8th century.

Qur'ān and *jihād*

Besides a measure of economic justice and the creation of a strong community ideal, the Prophet Muḥammad effected a general reform of the Arab society, in particular protecting its weaker segments—the poor, the orphans, women, and slaves. Slavery was not legally abolished, but emancipation of slaves was religiously encouraged as an act of merit. Slaves were given legal rights, including the right of acquiring their freedom against payment, in installments, of a sum agreed upon by the slave and his master out of his earnings. A slave woman who bore a child by her master became automatically free after her master's death. The infanticide of girls that was practiced among certain tribes—out of fear of poverty or a sense of shame—was forbidden.

Distinction and privileges based on tribal rank or race were repudiated in the Qur'ān and in the celebrated "Farewell Pilgrimage Address" of the Prophet shortly before his death. All men are therein declared to be "equal children of Adam," and the only distinction recognized in the sight of God is to be based on piety and good acts. The age-old Arab institution of intertribal revenge (called *tha'r*)—whereby it was not necessarily the killer who was executed but a person equal in rank to the slain person—was abolished. The pre-Islāmic ethical ideal of manliness was modified and replaced by a more humane ideal of moral virtue and piety.

(F.R./Ed.)

ORIGINS AND COMPILATION OF THE QUR'ĀN

Muslim tradition. According to Muslim tradition the Qur'ān was revealed to Muḥammad in separate pieces over some 20 years. On such occasions, Muḥammad, it is said, was in a kind of trance or ecstasy, during which the revelations were brought to him by the angel Gabriel. On his return to normal consciousness he recited the words of revelation to those present. There are many traditions about the occasions on which a certain *sūrah* or part of a *sūrah* was revealed. Thus the revelation of the Qur'ān is connected with events in the life of the Prophet. Even the traditional recension (version) of the Qur'ān itself classifies the *sūrah*s as Meccan or Medinan.

Revelation of the Qur'ān to the Prophet

Obviously, many people learned the words of the revelation by heart, but there are also traditions that, at the time of their revelation, Muḥammad had them written down on "pieces of paper, stones, palm-leaves, shoulder-blades, ribs, and bits of leather," *i.e.,* whatever writing-material there was at hand. It is believed that the Prophet indicated to the scribes the context in which a certain passage should be placed.

After the Prophet's death, and especially after the battle of Yamāmah (633), in which a great number of those who knew the Qur'ān by heart had fallen, fear arose that the knowledge of the Qur'ān might disappear. So it was decided to collect the revelations from all available written sources and, as Muslim tradition has it, "from the hearts [*i.e.,* memories] of people." A companion of the Prophet, Zayd ibn Thābit, is said to have copied on sheets whatever he could find and to have handed it over to the caliph 'Umar. After 'Umar's death the collection was left in the care of his daughter Ḥafṣah. Other copies of the Qur'ān appear to have been written later, and different versions were used in different parts of the Muslim empire. So that there would be no doubt about the correct reading of the Qur'ān, the caliph 'Uthmān (644–656) is reported to have commissioned Zayd ibn Thābit and some other learned men to revise the Qur'ān using the "sheets" of Ḥafṣah, comparing them with whatever material was at hand, and consulting those who knew the Qur'ān by heart. It was decided that in case of doubt about the pronunciation, the dialect of Quraysh, the Prophet's tribe, was to be given preference. Thus an authoritative text of the Qur'ān (now known as the 'uthmānic recension) was established.

Establishment of an authoritative text

These traditions may have been reworked and changed to some extent to suit certain dogmatic theories concerning the Qur'ān, but in the main they reflect historical truth. It is obvious that the description of the method of revelation has been somewhat simplified. The Qur'ān itself states (42:50–52) that God spoke to Muḥammad "by suggestion, or from behind a veil, or by sending a messenger to suggest what he pleases." The first term (Arabic *waḥy*) denotes a "suggestion" or "inspiration" of the kind that is well known by many poets; the Qur'ān also uses a term meaning "it was sent down." The second term seems to suggest some kind of imaginative locution without any accompanying vision. Only the third expression alludes to an angel but without mentioning the name of Gabriel.

Views of those outside Islām. The chronology of the *sūrah*s is a much debated problem. The existing traditions concerning the occasions for the revelation of certain passages cannot always be controlled and may or may not be reliable. European scholars have applied the criteria of style and contents to establish the relative order of the *sūrah*s or parts of *sūrah*s. From the time when Theodor Nöldeke published his *History of the Qur'ān* (1860), it has been common to arrange the *sūrah*s in four groups, deriving from three subsequent periods at Mecca and from Medina. The above exposition of the content of the Qur'ān roughly follows this arrangement.

In the Muslim view, Muḥammad received every word of the Qur'ān directly from God. The Qur'ān describes, and indignantly rejects, accusations that the Prophet had reproduced things that he had drawn from other sources. Western scholars who have analyzed the contents of the various revelations have shown that much of the narrative material concerning biblical persons and events differs from the biblical account and seems to have come from later Christian and, above all, from Jewish sources (*e.g.*, Midrash). Other motifs, such as the idea of the impending judgment and the descriptions of paradise agree with standard topics in the missionary preaching of the contemporary Syriac church fathers. The dependence need not, however, be of a literary kind, but might be due to influence from oral traditions.

Question of influence of oral traditions on text

It would appear that learning the words of the revelation by heart was the normal way of preserving them, and that only on special occasions were the words written down immediately. The existence of various early collections of Qur'ānic material seems to be a warranted fact, although their nature and contents cannot be determined. Some of the *sūrah*s beginning with separate letters (*al-fawātiḥ*)—certain consonant combinations detached from the main text (mentioned above under the heading *Form*)—occur together in the present Qur'ān and in the order of decreasing length in such a way as to suggest that they once formed separate collections. The establishment of a vulgate recension (a standard version) was not sufficient to secure the uniform and correct reading of the Qur'ān in all details. The Arabic script was incomplete; several consonants were easy to confuse, and there was no way of indicating the vowels to differentiate the variety of possible meanings inherent in a particular combination of consonants. To assure the correct recitation, therefore, it was necessary to know the text more or less by heart. In this way, differing variant readings arose, warranted by this or that "reader" of the Qur'ān.

The recorded variations, however, turned out to be remarkably few, and though no complete listing of the textual variants exists, it can safely be said that the textual tradition of the Qur'ān is much firmer and more uniform than that of the New Testament. The Arabic script was gradually improved. Diacritical signs were introduced to distinguish the letters that were similar in form, and long vowels were indicated by the letters *alif* (for *ā*), *wāw* (for *ū*), and *yā* (for *ī*). It is known that this vowel system was still disputed at the beginning of the 9th century. The special vowel signs placed above or beneath the letters were added in a different colour and did not count as part of the text itself.

INTERPRETATIONS

The "readers" (*qurrā'*, singular *qāri'*) were the specialists of the text of the Qur'ān. They were at the same time philologians, and it was to a great extent from their dealings with the language of the Qur'ān that the science of Arabic grammar grew. Two schools developed, one at Baṣra (in present-day Iraq), which was especially interested in systematizing and ordering the material to set up the rules governing the language, and a rival one at Kūfa (also in Iraq), which took more interest in the exceptional. It was theorized that several variant readings could be accepted only if they were based on the 'Uthmānic recension (version). It was also important that a reading be based on the authority of some renowned reader.

There was also theological speculation as to the true nature of the Qur'ān. In the discussions initiated by the Mu'tazilites (Seceders; literally, "those who stand apart"; a group that sought to introduce philosophical principles from Greek rationalism into Islāmic thought) the question of the eternity of the Qur'ān (*i.e.*, of its heavenly prototype) was one of the main points. The Mu'tazilites, who wanted to avoid everything that might compromise or encroach upon the oneness of God, denied the doctrine that the Qur'an was uncreated and eternal, because this would mean that something else besides the God of eternity would exist eternally and thus create an eternal and irreconcilable "dualism." Consequently they asserted that the Qur'ān was created by God. This doctrine, however, was rejected by orthodox adherents of Islām. In popular belief, the reverence for the Qur'ān is often directed toward the visible, physical book or parts of it. Oaths are taken on it, and passages are sometimes copied out of it to be used for magical or superstitious purposes.

In these and other doctrinal disputes the parties sought support for their opinions in the sayings of the Qur'ān, since it was considered as the ultimate authority in all legal and religious questions. The correct interpretation of the Qur'ān became the object of a special branch of learning, the so-called *tafsīr*, or Qur'ānic exegesis. All kinds of resources were utilized in order to elucidate the meaning of a Qur'ānic passage. Traditions concerning the circumstances surrounding the revelation of certain passages or containing interpretative utterances of the Prophet that had been transmitted orally were recorded and collected, together with other traditions deriving from and concerning the Prophet (Ḥadīth). At times, in order to provide authority for a certain theory, traditions were simply invented. Any interpretation of a Qur'ānic passage that could not be supported by Ḥadīth was originally rejected. The results of the study of grammar and lexicography were also utilized; examples from contemporary poetry were often quoted in order to elucidate the grammatical structure or the lexical meaning of a passage. Thus, work on the Qur'ān, whose ultimate goal was the correct understanding and application of its teachings, went hand in hand with the development of Arabic grammar and lexicography.

Qur'ānic exegesis

Two works are especially renowned in the field of *tafsīr*, namely the commentary of aṭ-Ṭabarī (839–923), a huge encyclopaedic collection that sums up everything that had been done so far in the field, and the *Kashshāf* of Zamakhsharī (1075–1143), which has gained almost canonical reputation, though its author was a Mu'tazilite and began his work with the words, "Praise be to God who created the Qur'ān." A handy commentary of Bayḍāwī (d. *c.* 1280), which is often quoted as authoritative, is merely an abridged revision of the latter work.

The theological schools of medieval Islām all sought to support their doctrines with the aid of Qur'ānic exegesis, and each of them produced their own commentaries. There are also examples of allegorical interpretation (*ta'wīl*) especially in Ṣūfī (Islāmic mystical) literature, in which the doctrines of mysticism are found to be hidden behind the literal sense of the Qur'ānic word.

Qur'ānic exegesis gained new significance with the appearance of modernism toward the end of the 19th century. The modernists, who sought to revive Islām from its degradation and to reconcile it with what they found valuable in Western scientific traditions, set up the principle of returning to the pure and uncorrupted Islām of the "ancestors." As a consequence, the interpretation of the oldest and original source of Islām was regarded as imperative, and attempts were made to establish the principles necessary for a correct understanding of the Qur'ān. Traditional exegesis was accused of having introduced Israelite legends and false traditions that had nothing to do with the original teachings of the Prophet. On the other hand, the authority of the Qur'ān was never called in question.

Muḥammad 'Abduh, the founder of modernism in Egypt, for several years published exegetical lectures in the journal *al-Manār;* and they were later published in book form by his Syrian disciple Rashīd Riḍā. In them he accepts the Qur'ān as the literally inspired word of God, in which there can be nothing false or antiquated, and tries to show that the results of modern science and many modern views are already present in the Qur'ān. This is often achieved by twisted interpretations, reading modern ideas into the words of the Qur'ān. For instance, the *jinn* (genii) of *sūrah* 2:176 that cause disease are interpreted as "microbes," and the words in 2:250, "How often a little company has overcome a numerous company; and God is with those who endure," is taken to refer to ideas reminiscent of Darwin's theory of the struggle for life and the survival of the fittest. Allegorical interpretation is also used when it can serve the purpose of the author. Other modernistic interpreters of the Qur'ān have continued along the same lines. The Qur'ān is, however, left untouched by criticism; as the infallible word of God it cannot have been influenced by the circumstances under which it was revealed, it can contain no mistake, and it cannot be superseded by any new discovery.

Modern commentaries

Later developments, however, have brought some new ideas to the fore. In an Urdu commentary on the Qur'ān, which has in part been made available in English, Maulana Abul Kalam Azad (1888–1958), an Indian Muslim scholar (minister of education of the Republic of India at the time of his death), developed some new principles for the interpretation of the Qur'ān. He argues that it is necessary to interpret the Qur'ān against the background of its environment; therefore it is necessary to study the cultures and the languages of ancient Arabia and other Semitic peoples. Study of the historical circumstances in which the Qur'ān came into being is said to facilitate the understanding of what it meant to those who received the revelation.

Scholars have no doubt, however, that there are new developments in the field of Qur'ānic exegesis. D. Rahbar, in his study *The God of Justice* (1960), argues that in order to elucidate a passage in the Qur'ān one should quote traditional exegesis and medieval dogmatics and, above all, use other Qur'ānic passages for comparison, letting one passage throw light on another. Though such ideas are looked upon with suspicion by orthodox Muslims and are fervently rejected by most Muslim leaders, they may indicate the inception of a more historical view of the Qur'ān, one that tries to distinguish between central religious ideas and those outward things that are dependent on the historical environment.

TRANSLATIONS

The Qur'ān was revealed to Muḥammad as "an Arabic book" or an Arabic reading (*qur'ān*), to provide the Arabs with a holy book in their own language, comparable with the Scriptures of Judaism and Christianity. As has been noted, the language of the Qur'ān is regarded as surpassing everything that can be written in Arabic. The Qur'ān itself is a miracle and cannot be imitated by man.

As a consequence of this, it is regarded as unfitting to translate the Qur'ān. In countries in which other languages are spoken, the Qur'ān is still recited in Arabic. There exist Muslim translations of the Qur'ān; *e.g.*, into Turkish, Urdu, and English (the latter during the Aḥmadiyah movement founded in 1889 by Mirza Ghulam Aḥmad in the Punjab region of India), but on principle these are regarded as paraphrases, not as translations that can be used for ritual purposes.

The Qur'ān was first printed in Arabic at Rome by Pagninus Brixiensis (1530), but the edition was never circulated. A. Hinckelmann published an Arabic text at Hamburg in 1694. Since then several European editions have appeared; one of the best was that of G. Flügel (1834), the first critical edition, often reprinted. It is from this edition that Western scholars have usually quoted the Qur'ān. Several editions are today printed in Muslim countries, and an official Egyptian edition is gaining more and more ground among Western scholars.

First critical edition

The first Latin translation was made in 1143 at the request of an abbot of the monastery of Cluny and was published at Basle in 1543 by Theodor Bibliander and afterward rendered into Italian, German, and Dutch. The first French translation was by A. du Ryer (1647); it was translated into English by Alexander Ross (1649–88). G. Sale's English translation first appeared in 1734 and has passed through many new editions. It has become something of a classic and can still be useful in many respects. A translation by J.M. Rodwell, with the *sūrah*s arranged in chronological order, appeared in 1861. E.H. Palmer's translation was published in Sacred Books of the East in 1880. Bell's translation "with a critical rearrangement of the sūrahs" (1937–39) tries to analyze the *sūrah*s into their smallest units and show how these were joined together to form the present Qur'ān. (See *Bibliography* for contemporary English translations.)

The Qur'ān has also been translated into most other European languages. Special mention should be made of R. Blachère's French translation (1949–50) because of its rather detailed notes, and of R. Paret's German rendering (1962), which is very accurate and makes extensive use of parallel passages within the Qur'ān itself, but is rather dry in its style.

(H.R./Ed.)

Ḥadīth, traditions of the Prophet

Ḥadīth is the record of the traditions or sayings of the Prophet Muḥammad, revered and received as a major source of religious law and moral guidance, second only to the authority of the Qur'ān, or scripture of Islām. It might be defined as the biography of Muḥammad perpetuated by the long memory of his community for their exemplification and obedience. The development of Ḥadīth is a vital element during the first three centuries of Islāmic history, and its study provides a broad index to the mind and ethos of Islām.

NATURE AND ORIGINS

The term Ḥadīth derives from the Arabic root *ḥdth*, meaning "to happen," and so, "to tell a happening," "to report," "to have, or give, as news," or "to speak of." It means tradition seen as narrative and record. From it comes *sunnah* (literally, a "well-trodden path," *i.e.*, taken as precedent and authority or directive), to which the faithful conform in submission to the sanction that Ḥadīth possesses and that legalists, on that ground, can enjoin. Tradition in Islām is thus both content and constraint, Ḥadīth as the biographical ground of law and *sunnah* as the system of obligation derived from it. In and through Ḥadīth, Muḥammad may be said to have shaped and determined from the grave the behaviour patterns of the household of Islām by the posthumous leadership his personality exercised. There were, broadly, two factors operating to this end. One was the unique status of Muḥammad in the genesis of Islām; the other was the rapid geographical expansion of the new faith in the first two centuries of its history into various areas of cultural confrontation. Ḥadīth cannot be rightly assessed unless the measure of these two elements and their interaction is properly taken.

Sunnah

The experience of Muslims in the conquered territories of west and middle Asia and of North Africa was related to their earlier tradition. Islāmic tradition was firmly grounded in the sense of Muḥammad's personal destiny as the Prophet—the instrument of the Qur'ān and the apostle of God. The clue to tradition as an institution in Islām may be seen in the recital of the *Shahādah* or "witness" ("There is no god but God; Muḥammad is the prophet of God"), with its twin items as inseparable convictions—God and the messenger. Islāmic tradition follows from the primary phenomenon of the Qur'ān, received personally by Muḥammad and thus inextricably bound up with his person and the agency of his vocation. Acknowledgment of the Qur'ān as scripture by the Islāmic community was inseparable from acknowledgment of Muḥammad as its appointed recipient. In that calling, he had neither fellow nor partner, for God, according to the Qur'ān, spoke only to Muḥammad. When Muḥammad died, therefore, in AD 632, the gap thus created in the emotions and the mental universe of Muslims was shatteringly wide. It was also permanent. Death had also terminated the revelation embodied in the Qur'ān. By the same stroke scriptural mediation had ended, as well as prophetic presence.

The Prophet's death was said to have coincided with the perfection of revelation. But the perfective closure of both the book and the Prophet's life, though in that sense triumphant, was also onerous, particularly in view of the new changing circumstances, both of space and time, in the geographical expansion of Islām. In all the new pressures of historical circumstance, where was direction to be sought? Where, if not from the same source as the scriptural mouthpiece, who by virtue of that consummated status had become the revelatory instrument of the divine word and could therefore be taken as an everlasting index to the divine counsel? The instinct for and the growth of tradition are thus integral elements in the very nature of Islām, Muḥammad, and the Qur'ān. Ongoing history and the extending dispersion of Muslim believers provided the occasion and spur for the compilation of Ḥadīth.

HISTORICAL DEVELOPMENT

The appeal of the ordered recollection of Muḥammad to the Islāmic mind did not become immediately formalized and sophisticated. On the contrary, there is evidence that the full development of Ḥadīth was slow and uneven. Time and distance had to play their role before memory became stylized and official.

Relation to early Arabic poetry

Literary tradition in pre-Islāmic Arabia. The first generation had its own immediacy of Islāmic experience, both within the life span of the Prophet and in the first quarter century afterward. It had also the familiar patterns of tribal chronicle in song and saga. Pre-Islāmic poetry celebrated the glory of each tribe and their warriors. Such poetry was recited in honour of each tribe's ancestors. The vigour and élan of original Islām took up these postures and baptized them into Muslim lore. The proud history of which Muḥammad was the crux was, naturally, the ardent theme, first of chronicle, and then of history writing. Both needed and stimulated the cherishing of tradition. The lawyers, in turn, took their clues from the same source. While the Qur'ān was being received, there had been reluctance and misgiving about recording the words and acts of the Prophet, lest they be confused with the uniquely constituted contents of the scripture. Knowledge of Muḥammad's disapproval of the practice of recording his words is evidence enough that the practice existed. With the Qur'ān complete and canonized, those considerations no longer obtained; and time and necessity turned the instinct for Ḥadīth into a process of gathering momentum.

Developments of the 1st and 2nd centuries AH. Within the first century of the Prophet's death, tradition had come to be a central factor in the development of law and the shape of society. Association by Ḥadīth with Muḥammad's name and example became increasingly the ground of authority. The 2nd century brought the further elaboration of this relationship by increasing formalism in its processes. Traditions had to be sustained by an expert "science" of attestation able to satisfy rigorous formal criteria of their connection with the person of Muḥammad through his "companions," by an unbroken sequence of "reportage" (see below). This science became so meticulous that it is fair (even if also paradoxical) to suspect that the more complete and formally satisfactory the attestation claimed to be, the more likely it was that the tradition was of late and deliberate origin. The developed requirements of acceptability that the tradition boasted simply did not exist in the early, more haphazard and spontaneous days.

It is clear that many customs and usages native to non-Arab societies prior to their Islāmization found their way into Islām in the form of reputed or alleged traditions of Muḥammad, though always on the condition of their general compatibility with the Islāmic religion. Implicit in this sense in Muḥammad's personal example and genius, tradition inferred an elasticity and an embrace large enough to comprehend and anticipate all that Islām in its wide geographical experience was to become.

Qur'ānic commentary

Qur'ānic commentary, as it developed in the wake of these other factors of law and custom, also leaned heavily on traditional material, for the incidents of the Qur'ānic narrative and the occasions of revelation could best be understood by what tradition had to say in its reporting of them. Further, since the patterns of Qur'ānic commentary were largely hortatory, Ḥadīth was a ready mine of word and story calculated to exemplify and reinforce what exhortation commended. Except in rare and controversial cases (the so-called Ḥadīth Qudsī, or Holy Tradition), these traditional factors in Qur'ānic interpretation were only elucidatory, and the substance of tradition could in no way dispute or displace the essential, primary, authority of the Qur'ānic text. For the *obiter dicta* (incidental observations) of Muḥammad, though sacrosanct, lacked the hallmark of revelation, which belonged solely to the Qur'ān. Among earliest developed examples of Ḥadīth are the narratives of the biographer Ibn Isḥāq (died AH 150 [AD 767]) and the compilation of laws by Mālik ibn Anas, known as al-Muwaṭṭa' (died AH 179 [AD 795]). But they preceded by less than half a century the success of the theory that made tradition indispensable to the valid development of Islāmic law.

3rd century AH and subsequent developments. The chief protagonist of the view correlating tradition and law was Muḥammad ash-Shāfi'ī (died AH 204 [AD 820]) who claimed for tradition a divine imprint as an extension of the revelation of the Qur'ān. It was in line with this conviction that the phrase "the Qur'ān and the *sunnah*" became current to describe the fount of authority in Sunnī Islām (the major traditionalist sect). By this mandate and out of the needs and inventiveness of lawyers, the mass of tradition grew apace. When virtually no issues could be argued, still less settled, except by connection with cited acts and opinions of Muḥammad, the temptation to require or to imagine or to allege such traditions became irresistible. Supply approximated to demand, and the growth of both made more ingenious and pretentious the science of supporting attribution. The increasing volume and complexity of the material contained in Ḥadīth necessitated larger compilations and more detailed classification. These factors worked together to inspire a critical editorial activity that in the course of the 3rd century generated what have come to be regarded as the six canonical collections of Ḥadīth by Sunnī Muslims. The first two of them have acquired a status of great sanctity. Before noting these it is convenient to describe the editorial task and the editorial procedures that constitute the developed science of Ḥadīth criticism.

THE SCIENCE OF ḤADĪTH

Matn and *isnād*

The study of tradition distinguishes between the substance, or content, known as the "gist" (*matn*) of the matter, and the "leaning" (*isnād*) or chain of corroboration on which it hangs.

Form of Ḥadīth and criteria of authentication. That Muḥammad observed, "Seek knowledge, though it be in China" or "Beware of suspicion, for it is the falsest of falsehoods" reveals the *matn* or "the meat of the matter." The formula introducing such a Ḥadīth would speak in the first person: "It was related to me by A, on the authority of B, on the authority of C, on the authority of D, from E (here a companion of Muḥammad) that the Prophet said . . . ." This chain of names constituted the *isnād* on which the saying or event depended for its authenticity. The major emphases in editing and arguing from tradition always fell on the *isnād*, rather than on a critical attitude to the *matn* itself. The question was not, "Is this the sort of thing Muḥammad might credibly be imagined to have said or done?" but "Is the report that he said or did it well supported in respect of witnesses and transmitters?" The first question would have introduced too great a danger of subjective judgment or independence of mind, though it may be suspected that issues were in fact often decided by such critical appraisal in the form of decisions ostensibly relating only to *isnād*. The second question certainly allowed a theoretically objective and reasonably precise pattern of criteria.

If the adjacent names in the chain of transmission overlapped in life, there was certainty that they could have listened to one another. Their travels were also investigated to see if their paths could have really crossed. Biographies could be built up to show that they were honest men and spoke truly. Comparative study could be made of their reputations for veracity as acknowledged by their contemporaries or indicated by their traditions when compared. The frequency of currency through several sources was yet another element in the testing of traditions. Most important of all was the final link with the "companion," who in the first instance had the tradition from his or her contact with the Prophet.

Weight of traditions

Classifications. In all these ways, and others involving more minutiae, it was possible to establish categories of Ḥadīth quality. Traditions might be sound (*ṣaḥīḥ*), good (*ḥasan*), or weak (*ḍa'īf*). Other terms, such as healthy (*ṣāliḥ*) and infirm (*saqīm*), were also current. Each of the three classifications was liable to subdivisions, depending on refinements of assessment and, later, on their standing with the classic compilers. Distinctions were less rigorously seen if the traditions were cited not for legal definitions but merely for moral purposes. A *ḍa'īf* tradition, for example, might well be salutary for exhortation, even if lawyers were required to exclude or ignore it. Traditions also varied in strength according to whether one or more "companions" could be adduced, whether the *isnād* had parallels, whether they were continuous back to Muḥammad (*muttaṣil*), or intermitted (*mawqūf*). The subtleties in these and other questions were part of the active competence that attended the whole science.

The repute and authority of the canonical collections did much to stabilize the situation, but only because their emergence demonstrated that the zest for tradition had overreached itself. By the end of the 3rd century AH it was sorely necessary to solidify Ḥadīth into a stable corpus of material to which no new element could credibly be added and from which extravagances had been purged. The Ḥadīth tradition within the various traditions had by then become a permanent and disciplined element in the authority structure of Islām—the second great source of law and practice, complementary to the Qur'ān and available for analogical handling (*qiyās*) and for consensus (*ijtihād*) as further sources of legislation, arguing from the Qur'ān and the Sunnah as primary. Shī'ah tradition (see below) stands apart from this structure of authority.

THE COMPILATIONS

The most revered of all traditionalists was Muḥammad ibn Ismā'īl al-Bukhārī (AH 194–56 [AD 810–870]), whose *Kitāb al-Jāmi' aṣ-Ṣaḥīḥ* (*The Book of the Authentic Collection*) has a unique place in the awe and esteem of Muslims as a work of great historical import and deep piety. While a boy he made the pilgrimage to Mecca and gathered traditions in wide travels. According to tradition, he was inspired to his task by a vision of Muḥammad pestered by flies while asleep—flies that he (al-Bukhārī) fanned from the Prophet's face. The flies represented the cloud of spurious traditions darkening the true image, and the fan was its tireless rescuer. Whatever the truth of this narrative, it captures the temper of al-Bukhārī's vocation. His *Ṣaḥīḥ* occupied 16 years of editorial pains and scrutiny. He included 7,397 traditions with full *isnād*. Allowing for repetitions, the net total was 2,762, gathered, it is said, from over 600,000 memorized items. He arranged the whole into 97 books and 3,450 chapters or topics, repeating the traditions that bore on several themes.

Of comparable stature was the *Ṣaḥīḥ* of Muslim ibn al-Ḥajjāj (AH 202–261 [AD 817–875]), to which the compiler prefaced a discussion of the criteria of Ḥadīth. The material largely confirms his contemporaries, and all such traditions common to these two authorities are known as agreed (*muttafaq*). It became characteristic to give freer rein to prevailing or communal assent in matters of *isnād*.

There are four other classical collections of tradition, all belonging within the 3rd century AH, and interdependent in part. Abū Dā'ūd al-Sijistānī (AH 202–275 [AD 817–889]) produced his *Kitāb as-Sunan* ("Book of traditions"), containing 4,800 traditions relating to matters of jurisprudence (as the term *Sunan* indicates, in contradistinction to a *Jāmi'*, or collection embracing all fields). Abū 'Isā Muḥammad at-Tirmidhī (died AH 279 [AD 892]) edited the *Jāmi'aṣ-Ṣaḥīḥ*, adding notes on the distinctive interpretations of the schools of law (*madhāhib*). Abū' Abd ar-Raḥmān an-Nasā'ī (AH 216–303 [AD 830–915]) produced another *Kitāb as-Sunan* with special concern for the religious law relating to ritual acts. Abū 'Abdallāh ibn Mājā (AH 210–273 [AD 824–886]), a pupil of Abū Dā'ūd, compiled another with the same title but tended to a readier tolerance of less than satisfactory traditions. Preferences shifted between these four, and some were slower of recognition than others. Nor did they oust the earlier collection of Mālik ibn Anas, which maintained, if intermittently, its wide appeal. But they formed the increasing reliance of generations of Muslims, within the unique eminence of the master "pair," and formed the sources of later popular editions, intended to conflate material for didactic purposes. One such was the work of Abū Muḥammad al-Baghawī (died AH 516 [AD 1122]) called *Maṣābiḥ as-Sunnah* ("The Lamps of the Sunnah"). Commentaries on all these classical *musannafāt*, or compilations, were many, and important in education and piety.

SECTARIAN VARIATIONS

The tradition of the Shī'ah, a minority branch of Islām, (distinguished from the tradition of the Sunnah majority by belief in the special role of the Prophet's cousin 'Alī and his descendants) diverges sharply from a very early date, though the emphasis on the personality of Muḥammad was identical. The Shī'ah broke away from the (to be) dominant Sunnī stream of Islām for deep reasons of politics, emotion, and theology. There was the dispute about caliphal succession and the role of 'Alī, cousin and son-in-law of Muḥammad and fourth caliph, and bitter cleavage because of the tragic fate of his two sons and especially of Ḥusayn in the massacre of Karbalā', from which there ultimately evolved the theology of vicarious suffering epitomized in Shī'ī devotion and ritual. (Sectarian disputes are treated in detail; see below *Theology and sectarianism*.) All these factors inevitably involved the business of tradition. The schism read the origins according to the divided loyalties, and there was little that was not potentially contentious, apart from obvious matters; *e.g.*, Muḥammad's intentions for 'Alī and the caliphate. The issues were fought out in rivalry for the mind of the Prophet, the authority of which was the sole agreement in the very disputing of it. The Shī'ah thus rejected the tradition of the Sunnīs and developed their own corpus of tradition (though there is evidence that an-Nasā'ī, at least, among the classical compilers, had sympathy with aspects of their cause). They also questioned the Sunnī notions of *isnād* and of the community as a locus of authority and evolved their own system of submission to their *imāms* (Shī'ah leaders). This altered the whole role that tradition might play. The major Shī'ī compilations date from the 4th and 5th centuries and allow only traditions emanating from the house of 'Alī. The first of them is that of Abū Ja'far Muḥammad al-Qulīnī (died AH 328 [AD 939]), *Kāfī fī 'Ilm ad-Dīn*, which might be translated: "All You Need About the Science of Religious Practice."

Shī'ah and Sunnah

SIGNIFICANCE OF ḤADĪTH

Canonical collections of Ḥadīth are, for the non-Muslim, an introduction to a world of faith, of behaviour and authority, a world of almost encyclopaedic inclusiveness. Provisions of law are the primary element, enlarging Qur'ānic legislation. They contain a whole array of moral, social, commercial, and personal matters, as well as the themes of eschatology. All reaches of public and private conduct may be found there, from the disposal of a date stone to the crisis of the deathbed, from the manner of ablution to the duties of forgiveness, from the physical routines of digestion to the description of the day of judgment. There is a Talmudic capacity for detail and scrupulousness in legal and ethical prescriptions and precepts. There are stories of integrity and right action, for example, that of the purchaser of a plot of ground who subsequently unearthed in it a pot of gold, which he brought back to the former owner, protesting that it was not within his bargain. The vendor, likewise, refused to claim it since he had not known the gold was there when he sold his field. An arbitrator solved their dilemma of honesty by proposing the marriage of the son of one with the daughter of the other so that, after alms, the gold might be settled on the couple. Through and in tradition, Islām aligned itself authoritatively with all it found compatible in local usages and brought hospitably and masterfully within its purview the continuity of many cultures. There is wide evidence of the impact of Jewish and Christian elements, notably in the realm of eschatology, in the elaboration of the stark and urgent Qur'ānic doctrine of the last judgment. But always the imprint of Islām is clear. Tradition is at once a mine and a kind of currency, the source and the circulation of the values it makes and preserves. (A.K.C./Ed.)

Fundamental practices and institutions of Islām

THE FIVE PILLARS

During the earliest decades after the death of the Prophet, certain basic features of the religio-social organization of Islām were singled out to serve as anchoring points of the community's life and formulated as the "Pillars of Islām." To these five, the Khawārij sect added a sixth pillar, the *jihād*, which, however, was not accepted by the general community.

The shahādah, or profession of faith. The first pillar is the profession of faith: "There is no god but God; Muḥammad is the prophet of God," upon which depends the membership in the community. The profession of faith must be recited at least once in one's lifetime, aloud, correctly, and purposively, with an understanding of its meaning and with an assent from the heart. From this fundamental belief are derived beliefs in (1) angels (particularly Gabriel, the Angel of Revelation), (2) the revealed Books (the Qur'ān and the sacred books of Judeo-Christian revelation described in the Qur'ān), (3) a series of prophets (among whom figures of the Judeo-Christian tradition are particularly eminent—although it is believed that God has sent messengers to every nation), and (4) the Last Day (Day of Judgment).

Basic beliefs deriving from the *shahādah*

Prayer. The second pillar consists of five daily congregational prayers. These prayers may be offered individually if one is unable to go to the mosque. The first prayer is performed before sunrise, the second just after noon, the third in the later afternoon, the fourth immediately after sunset, and the fifth before retiring to bed.

Before a prayer, ablutions, including the washing of hands, face, and feet, are performed. The muezzin (one who gives the call for prayer) chants aloud from a raised place (such as a tower) in the mosque. When prayer starts, the *imām,* or leader (of the prayer), stands in the front facing in the direction of Mecca, and the congregation stands behind him in rows, following him in various postures. Each prayer consists of two to four genuflection units (*rakʿah*); each unit consists of a standing posture (during which verses from the Qur'ān are recited—in certain prayers aloud, in others silently), as well as a genuflection and two prostrations. At every change in posture, "God is great" is recited. Tradition has fixed the materials to be recited in each posture.

Special congregational prayers are offered on Friday instead of the prayer just after noon. The Friday service consists of a sermon (*khuṭbah*), part of which consists of preaching in the local language and part of recitation of certain formulas in Arabic. In the sermon, the preacher usually recites a verse of the Qur'ān and builds his address on it, which can be of a moral, social, or political content. Friday sermons have usually considerable impact on public opinion regarding sociopolitical questions.

Nature of sermons

Although not ordained as an obligatory duty, nocturnal prayers (called *tahajjud*) are encouraged, particularly during the latter half of the night. During the month of Ramaḍān (see below *Fasting*), lengthy prayers are offered congregationally before retiring and are called *tarāwīḥ.*

In strict doctrine, the five daily prayers cannot be waived even for the sick, who may pray in bed and, if necessary, lying down. When on a journey, the two afternoon prayers may be combined into one; the sunset and late evening prayers may be combined as well. In practice, however, much laxity has occurred, particularly in modern times, although Friday prayers are still well attended.

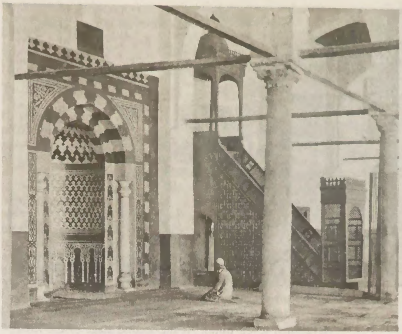
The zakāt. The third pillar is the obligatory tax called *zakāt* ("purification," indicating that such a payment makes the rest of one's wealth religiously and legally pure). This is the only permanent tax levied by the Qur'ān and is payable annually on food grains, cattle, and cash after one year's possession. The amount varies for different categories. Thus, on grains and fruits it is 10 percent if land is watered by rain, 5 percent if land is watered artificially. On cash and precious metals it is 2½ percent. *Zakāt* is collectable by the state and is to be used primarily for the poor, but the Qur'ān mentions other purposes: ransoming Muslim war captives, redeeming chronic debts, paying tax collectors' fees, *jihād* (and by extension, according to Qur'ān commentators, education and health), and creating facilities for travellers.

After the breakup of Muslim religio-political power, payment of *zakāt* has become a matter of voluntary charity dependent on individual conscience. Some Muslim countries are seeking to reintroduce it, and in several Middle Eastern countries *zakāt* is officially collected, but on a voluntary basis.

The fast of Ramaḍān

Fasting. Fasting during the month of Ramaḍān (ninth month of the Muslim lunar calendar), laid down in the Qur'ān (2:183–185), is the fourth pillar of the faith. Fasting begins at daybreak and ends at sunset, and during the day eating, drinking, and smoking are forbidden. The Qur'ān (2:185) states that it was in the month of Ramaḍān that the Qur'ān was revealed. Another verse of the Qur'ān (97:1) states that it was revealed "on the night of determination," which Muslims generally observe on the night of 26–27 Ramaḍān. For a person who is sick or on a journey, fasting may be postponed until "another equal number of days." The elderly and the incurably sick are exempted through the daily feeding of one poor person.

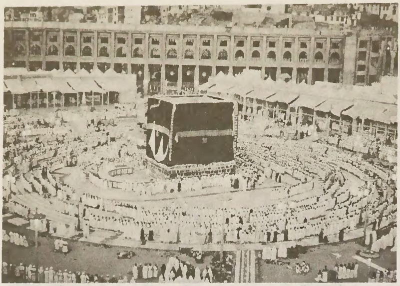
The ḥajj. The fifth pillar is the annual pilgrimage (*ḥajj*) to Mecca prescribed for every Muslim once in a lifetime—"provided one can afford it" and provided a person has enough provisions to leave for his family in his absence. A special service is held in the Sacred Mosque on the 7th of the month of Dhū al-Ḥijjah (last in the Muslim year). Pilgrimage activities begin by the 8th and conclude on the 12th or 13th. All worshippers enter the state of *iḥrām;* they wear two seamless garments and avoid sexual intercourse, the cutting of hair and nails, and certain other activities. Pilgrims from outside Mecca assume *iḥrām* at specified points en route to the city. The principal activities consist of walking seven times around the Kaʿbah, a shrine within the mosque; the kissing and touching of the Black Stone (Ḥajar al-Aswad); and the ascent of and running between Mt. Ṣafā and Mt. Marwah (which are now, however, mere



Interior of the Mosque of ʿAmr ibn al-ʿĀṣ, Cairo, showing the *miḥrāb* (prayer niche) and the *minbar* (pulpit).
Lehnert & Landrock

Brian Brake—Magnum

Muslims at prayer, Kashmir, India.



Worshippers encircling the holy Kaʿbah, Mecca, Saudi Arabia.
By courtesy of the Saudi Arabian Information Service, Royal Embassy of Saudi Arabia

elevations) seven times. At the second stage of the ritual, the pilgrim proceeds from Mecca to Minā, a few miles away; from there he goes to ʿArafāt, where it is essential to hear a sermon and to spend one afternoon. The last rites consist of spending the night at Muzdalifah (between ʿArafāt and Minā) and offering sacrifice on the last day of *iḥrām*, which is the *ʿīd* ("festival") of sacrifice.

Many countries have imposed restrictions on the number of outgoing pilgrims because of foreign-exchange difficulties. Because of the improvement of communications, however, the total number of visitors has greatly increased in recent years. By the early 1990s the number of visitors was estimated to be about 2,000,000, approximately half of them from non-Arab countries. All Muslim countries send official delegations on the occasion, which is being increasingly used for religio-political congresses. At other times in the year, it is considered meritorious to perform the lesser pilgrimage (*ʿumrah*), which is not, however, a substitute for the *ḥajj* pilgrimage.

SACRED PLACES AND DAYS

The most sacred place for Muslims is the Kaʿbah sanctuary at Mecca, the object of the annual pilgrimage. It is much more than a mosque; it is believed to be the place where the heavenly bliss and power touches the earth directly. According to Muslim tradition, the Kaʿbah was built by Abraham. The Prophet's mosque in Medina is the next in sanctity. Jerusalem follows in third place in sanctity as the first *qiblah* (*i.e.*, direction in which the Muslims offered prayers at first, before the *qiblah* was changed to the Kaʿbah) and as the place from where Muḥammad, according to tradition, made his ascent (*miʿrāj*) to heaven. For the Shīʿah, Karbalāʾ in Iraq (the place of martyrdom of ʿAlī's son, Ḥusayn) and Meshed in Iran (where Imām ʿAlī ar-Riḍā is buried) constitute places of special veneration where the Shīʿah make pilgrimages.

Shrines of Ṣūfī saints. For the Muslim masses in general, shrines of Ṣūfī saints are particular objects of reverence and even veneration. In Baghdad the tomb of the greatest saint of all, ʿAbd al-Qādir al-Jīlānī, is visited every year by large numbers of pilgrims from all over the Muslim world.

The Ṣūfī shrines, which were managed privately in earlier periods, are almost entirely owned by governments in the late 20th century and are managed by departments of *awqāf* (plural of *waqf*, a religious endowment). The official appointed to care for a shrine is usually called a *mutawallī*. In Turkey, where such endowments formerly constituted a very considerable portion of the national wealth, all were confiscated by the regime of Atatürk (president, 1928–38).

The mosque. The general religious life of the Muslims is centred around the mosque, and in the days of the Prophet and early caliphs the mosque was, indeed, the centre of all community life. Small mosques are usually supervised by the *imām* (one who administers the prayer service) himself, although sometimes also a muezzin is appointed. In larger mosques, where Friday prayers are offered, a *khaṭīb* (one who gives the *khuṭbah*, or sermon) is appointed for Friday service. Many large mosques also function as religious schools and colleges. Mosque officials are appointed by the government in most countries. In some countries—*e.g.*, Pakistan—most mosques are private and are run by the local community, although some of the larger ones are being increasingly taken over by the government departments of *awqāf*.

Holy days. The Muslim calendar (based on the lunar year) dates from the emigration (*hijrah*) of the Prophet from Mecca to Medina in AD 622. The two festive days in the year are the *ʿīds*, ʿĪd al-Fiṭr celebrating the end of the month of Ramaḍān and the other, ʿĪd al-Aḍḥā (the feast of sacrifice), marking the end of the pilgrimage. Because of the crowds, *ʿīd* prayers are offered either in very large mosques or on specially consecrated grounds. Other sacred times include the "night of determination" (believed to be the night in which God makes decisions about the destiny of individuals and the world as a whole) and the night of the ascension of the Prophet to heaven. The Shīʿah celebrate the 10th of Muḥarram (the first month of the Muslim year) to mark the day of the martyrdom of Ḥusayn. The Muslim masses also celebrate the death anniversaries of various saints in a ceremony called *ʿurs* (literally, "nuptial ceremony"). The saints, far from dying, are believed to reach the zenith of their spiritual life on this occasion. (F.R./Ed.)

ISLĀMIC THOUGHT

Islāmic theology (*kalām*) and philosophy (*falsafah*) are two traditions of learning developed by Muslim thinkers who were engaged, on the one hand, in the rational clarification and defense of the principles of the Islāmic religion (*mutakallimūn*) and, on the other, in the pursuit of the ancient (Greek and Hellenistic, or Greco-Roman) sciences (*falāsifah*). These thinkers took a position that was intermediate between the traditionalists, who remained attached to the literal expressions of the primary sources of Islāmic doctrines (the Qurʾān, or the Islāmic scripture, and the Ḥadīth, or the sayings and traditions of Muḥammad) and who abhorred reasoning, and those whose reasoning led them to abandon the Islāmic community (the *ummah*) altogether. The status of the believer in Islām remained in practice a juridical question, not a matter for theologians or philosophers to decide. Except in regard to the fundamental questions of the existence of God, Islāmic revelation, and future reward and punishment, the juridical conditions for declaring someone an unbeliever or beyond the pale of Islām were so demanding as to make it almost impossible to make a valid declaration of this sort about a professing Muslim. In the course of events in Islāmic history, representatives of certain theological movements, who happened to be jurists and who succeeded in converting rulers to their cause, made those rulers declare in favour of their movements and even encouraged them to persecute their opponents. Thus there arose in some localities and periods a semblance of an official, or orthodox, doctrine.

Origins, nature, and significance of Islāmic theology

EARLY DEVELOPMENTS

Beginnings of Islāmic theology

The beginnings of theology in the Islāmic tradition in the second half of the 7th century are not easily distinguishable from the beginnings of a number of other disciplines—Arabic philology, Qurʾānic interpretation, the collection of the sayings and deeds of the prophet Muḥammad (Ḥadīth), jurisprudence, and historiography. Together with these other disciplines, Islāmic theology is concerned with ascertaining the facts and context of the Islāmic revelation and with understanding its meaning and implications as to what Muslims should believe and do after the revelation had ceased and the Islāmic community had to chart its own way. During the first half of the 8th century, a number of questions—which centred on God's unity, justice, and other attributes and which were relevant to man's freedom, actions, and fate in the hereafter—formed the core of a more specialized discipline, which was called *kalām* ("speech"). This term (*kalām*) was used to designate the more specialized discipline because of the rhetorical and dialectical "speech" used in formulating the principal matters of Islāmic belief, debating them, and defending them against Muslim and non-Muslim opponents. Gradually, *kalām* came to include all matters directly or indirectly relevant to the establishment and definition of religious beliefs, and it developed its own necessary or useful systematic rational arguments about human knowledge and the makeup of the world. Despite various efforts by later thinkers to fuse the problems of *kalām* with those of philosophy (and mysticism), theology preserved its relative independence from philosophy and other nonreligious sciences. It remained true to its original traditional and religious point of view, confined itself within the limits of the Islāmic revelation, and assumed that these limits as it understood them were identical with the limits of truth.

THE HELLENISTIC LEGACY

Relationships to other religious communities

The pre-Islāmic and non-Islāmic legacy with which early Islāmic theology came into contact included almost all the religious thought that had survived and was being defended or disputed in Egypt, Syria, Iran, and India. It was transmitted by learned representatives of various Christian, Jewish, Manichaean (members of a dualistic religion founded by Mani, an Iranian prophet, in the 3rd century), Zoroastrian (members of a monotheistic, but later dualistic, religion founded by Zoroaster, a 7th-century-BC Iranian prophet), Indian (Hindu and Buddhist, primarily), and Ṣābian (star worshippers of Harran often confused with the Mandaeans) communities and by early converts to Islām conversant with the teachings, sacred writings, and doctrinal history of the religions of these areas. At first, access to this legacy was primarily through conversations and disputations with such men, rather than through full and accurate translations of sacred texts or theological and philosophic writings, although some translations from Pahlavi (a Middle Persian dialect), Syriac, and Greek must also have been available.

The characteristic approach of early Islāmic theology to non-Muslim literature was through oral disputations, the starting points of which were the statements presented or defended (orally) by the opponents. Oral disputation continued to be used in theology for centuries, and most theological writings reproduce or imitate that form. From such oral and written disputations, writers on religions and sects collected much of their information about non-Muslim sects. Much of Hellenistic (post-3rd century BC Greek cultural), Iranian, and Indian religious thought was thus encountered in an informal and indirect manner.

From the 9th century onward, theologians had access to an increasingly larger body of translated texts, but by then they had taken most of their basic positions. They made a selective use of the translation literature, ignoring most of what was not useful to them until the mystical theologian al-Ghazālī (flourished 11th–12th centuries) showed them the way to study it, distinguish between the harmless and harmful doctrines contained in it, and refute the latter. By this time Islāmic theology had coined a vast number of technical terms, and theologians (*e.g.*, al-Jāḥiẓ) had forged Arabic into a versatile language of science; Arabic philology had matured; and the religious sciences (jurisprudence, the study of the Qur'ān, Ḥadīth, criticism, and history) had developed complex techniques of textual study and interpretation. The 9th-century translators availed themselves of these advances to meet the needs of patrons. Apart from demands for medical and mathematical works, the translation of Greek learning was fostered by the early ʿAbbāsid caliphs (8th–9th centuries) and their viziers as additional weapons (the primary weapon was theology itself) against the threat of Manichaeanism and other subversive ideas that went under the name *zandaqah* ("heresy" or "atheism"). (M.S.M./Ed.)

Theology and sectarianism

Despite the notion of a unified and consolidated community, as taught by the Prophet, serious differences arose within the Muslim community immediately after his death. According to the Sunnah, or traditionalist faction—who now constitute the majority of Islām—the Prophet had designated no successor. Thus the Muslims at Medina decided to elect a separate chief. Because he would not have been accepted by the Quraysh, the *ummah*, or Muslim community, would have disintegrated. Therefore, two of Muḥammad's fathers-in-law, who were highly respected early converts as well as trusted lieutenants, prevailed upon the Medinans to elect a single leader, and the choice fell upon Abū Bakr, father of the Prophet's favoured wife, ʿĀʾ ishah. All of this occurred before the Prophet's burial (under the floor of ʿĀʾishah's hut, alongside the courtyard of the mosque).

According to the Shīʿah, or "Partisans" of ʿAlī, the Prophet had designated as his successor his son-in-law ʿAlī ibn Abī Ṭālib, husband of his daughter Fāṭimah and father of his only surviving grandsons, Ḥasan and Ḥusayn. His preference was general knowledge; yet, while ʿAlī and the Prophet's closest kinsmen were preparing the body for burial, Abū Bakr, ʿUmar, and Abū ʿUbaydah from Muḥammad's Companions in the Quraysh tribe, met with the leaders of the Medinans and agreed to elect the aging Abū Bakr as the successor (*khalīfah*, hence "caliph") of the Prophet. ʿAlī and his kinsmen were dismayed but agreed for the sake of unity to accept the *fait accompli* because ʿAlī was still young

After the murder of ʿUthmān, the third caliph, ʿAlī was invited by the Muslims at Medina to accept the caliphate. Thus ʿAli became the fourth caliph (656–661), but the disagreement over his right of succession brought about a major schism in Islām, between the Shīʿah, or "legitimists"—those loyal to ʿAlī—and the Sunnah, or "traditionalists." Athough their differences were in the first instance political, arising out of the question of leadership, theological differences developed over time.

THE KHAWĀRIJ

During the reign of the third caliph, ʿUthmān, certain rebellious groups accused the Caliph of nepotism and misrule, and the resulting discontent led to his assassination. The rebels then recognized the Prophet's cousin and son-in-law, ʿAlī, as ruler but later deserted him and fought against him, accusing him of having committed a grave sin in submitting his claim to the caliphate to arbitration. The word *khāraju*, from which *khārijī* is derived, means "to withdraw" and Khawārij were, therefore, seceders who believed in active dissent or rebellion against a state of affairs they considered to be gravely impious.

The basic doctrine of the Khawārij was that a person or a group who committed a grave error or sin and did not sincerely repent ceased to be Muslim. Mere profession of the faith—"there is no god but God; Muḥammad is the prophet of God"—did not make a person a Muslim unless this faith was accompanied by righteous deeds. In other words, good works were an integral part of faith and not extraneous to it. The second principle that flowed from their aggressive idealism was militancy, or *jihād*, which the Khawārij considered to be among the cardinal principles, or pillars, of Islām. Contrary to the orthodox view, they interpreted the Qurʾānic command about "enjoining good and forbidding evil" to mean the vindication of truth through the sword. The placing of these two principles together made the Khawārij highly inflammable fanatics, intolerant of almost any established political authority. They incessantly resorted to rebellion and as a result were virtually wiped out during the first two centuries of Islām.

Because the Khawārij believed that the basis of rule was righteous character and piety alone, any Muslim, irrespective of race, colour, and sex, could, in their view, become ruler—provided he or she satisfied the conditions of piety. This was in contrast to the claims of the Shīʿah (the party of Muḥammad's son-in-law, ʿAlī) that the ruler must belong to the family of the Prophet and to the doctrine of the Sunnah (followers of the Prophet's way) that the head of state must belong to the Prophet's tribe, *i.e.*, the Quraysh.

A moderate group of the Khawārij, the Ibāḍīs, avoided extinction, and its members are to be found today in North Africa and in Oman and other parts of East Africa, including Zanzibar Island. The Ibāḍīs do not believe in aggressive methods and, throughout medieval Islām, remained dormant. Because of the interest of 20th-century Western scholars in this sect, the Ibāḍīs have become active and have begun to publish their classical writings and their own journals.

The permanent influence of the Khawārij

Although Khārijism is now essentially a story of the past, it has left a permanent influence on Islām, because of reaction against it. It forced the religious leadership of the community to formulate a bulwark against religious intolerance and fanaticism. Positively, it has influenced the reform movements that have sprung up in Islām from time to time and that have treated spiritual and moral placidity and status quo with a quasi-Khawārij zeal and militancy.

THE MUʿTAZILAH

The question of whether works are an integral part of faith or independent of it, as raised by the Khawārij, led to another important theological question: are human acts the result of a free human choice, or are they predetermined by God? This question brought with it a whole series of questions about the nature of God and of man. Although the initial impetus to theological thought, in the case of the Khawārij, had come from within Islām, full-scale religious speculation resulted from the contact and confrontation of Muslims with other cultures and systems of thought.

As a consequence of translations of Greek philosophical and scientific works into Arabic during the 8th and 9th centuries and the controversies of Muslims with Dualists (*e.g.*, Gnostics and Manichaeans), Buddhists, and Christians, a more powerful movement of rational theology emerged; its representatives are called the Muʿtazilah (literally "those who stand apart," a reference to the fact that they dissociated themselves from extreme views of faith and infidelity). On the question of the relationship of faith to works, the Muʿtazilah—who called themselves "champions of God's unity and justice"—taught, like the Khawārij, that works were an essential part of faith but that a person guilty of a grave sin, unless he repented, was neither a Muslim nor yet a non-Muslim but occupied a "middle ground." They further defended the position, as a central part of their doctrine, that man was free to choose and act and was, therefore, responsible for his actions. Divine predestination of human acts, they held, was incompatible with God's justice and human responsibility. The Muʿtazilah, therefore, recognized two powers, or actors, in the universe—God in the realm of nature and man in the domain of moral human action. The Muʿtazilah explained away the apparently predeterministic verses of the Qurʾān as being metaphors and exhortations.

Emphasis on reason

They claimed that human reason, independent of revelation, was capable of discovering what is good and what is evil, although revelation corroborated the findings of reason. Man is, therefore, under moral obligation to do the right even if there were no prophets and no divine revelation. Revelation has to be interpreted, therefore, in conformity with the dictates of rational ethics. Yet revelation is neither redundant nor passive. Its function is twofold. First, its aim is to aid man in choosing the right, because in the conflict between good and evil man often falters and makes the wrong choice against his rational judgment. God, therefore, must send prophets, for he must do the best for man; otherwise, the demands of divine grace and mercy cannot be fulfilled. Secondly, revelation is also necessary to communicate the positive obligations of religion—*e.g.*, prayers and fasting—which cannot be known without revelation.

God is viewed by the Muʿtazilah as pure Essence, without eternal attributes, because they hold that the assumption of eternal attributes in conjunction with Essence will result in a belief in multiple coeternals and violate the pure, unadulterated unity of God. God knows, wills, and acts by virtue of his Essence and not through attributes of knowledge, will, and power. Nor does he have an eternal attribute of speech, of which the Qurʾān and other earlier revelations were effects; the Qurʾān was, therefore, created in time and was not eternal.

The promises of reward that God has made in the Qurʾān to righteous people and the threats of punishment he has issued to evildoers must be carried out by him on the Day of Judgment. For promises and threats are viewed as reports about the future, and if not fulfilled exactly those reports will turn into lies, which are inconceivable of God. Also, if God were to withhold punishment for evil and forgive it, this would be as unjust as withholding reward for righteousness. There can be neither undeserved punishment nor undeserved reward; otherwise, good may just as well turn into evil and evil into good. From this position it follows that there can be no intercession on behalf of sinners.

When, in the early 9th century, the ʿAbbāsid caliph al-Maʾmūn raised Muʿtazilism to the status of the state creed, the Muʿtazilite rationalists showed themselves to be illiberal and persecuted their opponents. Aḥmad ibn Ḥanbal (died 855), an eminent orthodox figure and founder of one of the four orthodox schools of Islāmic law, was subjected to flogging and imprisonment for his refusal to subscribe to the doctrine that the Qurʾān, the word of God, was created in time.

THE SUNNAH

In the 10th century a reaction began against the Muʿtazilah that culminated in the formulation and subsequent general acceptance of another set of theological propositions, which became Sunnī, or "orthodox" theology.

The issues raised by these early schisms and the positions adopted by them enabled the Sunnī orthodoxy to define its own doctrinal positions in turn. Much of the content of Sunnī theology was, therefore, supplied by its reactions to those schisms. The term *sunnah,* which means a "well-trodden path" and in the religious terminology of Islām normally signifies "the example set by the Prophet," in the present context simply means the traditional and well-defined way. In this context, the term *sunnah* usually is accompanied by the appendage "the consolidated majority" (*al-jamāʿah*). The term clearly indicates that the traditional way is the way of the consolidated majority of the community as against peripheral or "wayward" positions of sectarians, who by definition must be erroneous.

The way of the majority. With the rise of the orthodoxy, then, the foremost and elemental factor that came to be emphasized was the notion of the majority of the community. The concept of the community so vigorously pronounced by the earliest doctrine of the Qurʾān gained both a new emphasis and a fresh context with the rise of Sunnism. Whereas the Qurʾān had marked out the Muslim community from other communities, Sunnism now emphasized the views and customs of the majority of the community in contradistinction to peripheral groups. An abundance of tradition (Ḥadīth) came to be attributed to the Prophet to the effect that Muslims must follow the majority's way, that minority groups are all doomed to hell, and that God's protective hand is always on (the majority of) the community, which can never be in error. Under the impact of the new Ḥadīth, the community, which had been charged by the Qurʾān with a mission and commanded to accept a challenge, now became transformed into a privileged one that was endowed with infallibility.

Tolerance of diversity. At the same time, while condemning schisms and branding dissent as heretical, Sunnism developed the opposite trend of accommodation, catholicity, and synthesis. A putative tradition of the Prophet that says "differences of opinion among my community are a blessing" was given wide currency. This principle of toleration ultimately made it possible for diverse sects and schools of thought—notwithstanding a wide range of difference in belief and practice—to recognize and coexist with each other. No group may be excluded from the community unless it itself formally renounces Islām. As for individuals, tests of heresy may be applied to their beliefs, but, unless a person is found to flagrantly violate or deny the unity of God or expressly negate the prophethood of Muḥammad, such tests usually have no serious consequences. Catholicity was orthodoxy's answer to the intolerance and secessionism of the Khawārij and the severity of the Muʿtazilah. As a consequence, a formula was adopted in which good works were recognized as enhancing the quality of faith but not as entering into the definition and essential nature of faith. This broad formula saved the integrity of the community at the expense of moral strictness and doctrinal uniformity.

On the question of free will, Sunnī orthodoxy attempted a synthesis between man's responsibility and God's omnipotence. The champions of orthodoxy accused the Muʿtazilah of quasi-Magian Dualism (Zoroastrianism) insofar as the Muʿtazilah admitted two independent and original actors in the universe: God and man. To the orthodox it seemed blasphemous to hold that man could act wholly outside the sphere of divine omnipotence, which had been so vividly portrayed by the Qurʾān but which the Muʿtazilah had endeavoured to explain away in order to make room for man's free and independent action.

Influence of Al-Ashʿarī and al-Māturīdī. The Sunnī formulation, however, as presented by al-Ashʿarī and al-Māturīdī, Sunnī's two main representatives in the 10th century, shows palpable differences despite basic uniformity. Al-Ashʿarī taught that human acts were created by God and acquired by man and that human responsibility depended on this acquisition. He denied, however, that man could be described as an actor in a real sense. Al-Māturīdī, on the other hand, held that although God is the sole Creator of everything, including human acts, nevertheless, man is an actor in the real sense, for acting and creating were two different types of activity involving different aspects of the same human act.

Formulators of Sunnī doctrine

In conformity with their positions, al-Ashʿarī believed that man did not have the power to act before he actually acted and that God created this power in him at the time of action; and al-Māturīdī taught that before the action man has a certain general power for action but that this power becomes specific to a particular action only when the action is performed, because, after full and specific power comes into existence, action cannot be delayed.

Al-Ashʿarī and his school also held that human reason was incapable of discovering good and evil and that acts became endowed with good or evil qualities through God's declaring them to be such. Because man in his natural state regards his own self-interest as good and that which thwarts his interests as bad, natural human reason is unreliable. Independently of revelation, therefore, murder would not be bad nor the saving of life good. Furthermore, because God's Will makes acts good or bad, one cannot ask for reasons behind the divine law, which must be simply accepted. Al-Māturīdī takes an opposite position, not materially different from that of the Muʿtazilah: human reason is capable of finding out good and evil, and revelation aids human reason against the sway of human passions.

Despite these important initial differences between the two main Sunnī schools of thought, the doctrines of al-Māturīdī became submerged in course of time under the expanding popularity of the Ashʿarite school, which gained wide currency particularly after the 11th century because of the influential activity of the Ṣūfī theologian al-Ghazālī. Because these later theologians placed increasing emphasis on divine omnipotence at the expense of the freedom and efficacy of the human will, a deterministic outlook on life became characteristic of Sunnī Islām—reinvigorated by the Ṣūfī world view, which taught that nothing exists except God, whose being is the only real being. This general deterministic outlook produced, in turn, a severe reformist reaction in the teachings of Ibn Taymīyah, a 14th-century theologian who sought to rehabilitate human freedom and responsibility and whose influence has been strongly felt through the reform movements in the Muslim world since the 18th century.

THE SHĪʿAH

The Shīʿah are the only important surviving sect in Islām. As noted above, they owe their origin to the hostility between ʿAlī (the fourth caliph and son-in-law of the Prophet) and the Umayyad dynasty (661–750). After ʿAlī's death, the Shīʿah (Party; *i.e.*, of ʿAlī) demanded the restoration of rule to ʿAlī's family, and from that demand developed the Shīʿite legitimism, or the divine right of the holy family to rule. In the early stages, the Shīʿah used this legitimism to cover the protest against the Arab hegemony under the Umayyads and to agitate for social reform.

Gradually, however, Shīʿism developed a theological content for its political stand. Probably under Gnostic (esoteric, dualistic, and speculative) and old Iranian (dualistic) influences, the figure of the political ruler, the *imām* (exemplary "leader"), was transformed into a metaphysical being, a manifestation of God and the primordial light that sustains the universe and bestows true knowledge on man. Through the *imām* alone the hidden and true meaning of the Qurʾānic revelation can be known, because the *imām* alone is infallible. The Shīʿah thus developed a doctrine of esoteric knowledge that was adopted also, in a modified form, by the Ṣūfīs, or Islāmic mystics (see below *Islāmic mysticism, Ṣūfism*). The orthodox Shīʿah recognize 12 such *imāms*, the last (Muḥammad) having disappeared in the 9th century. Since that time, the *mujtahids* (*i.e.*, the Shīʿī divines) have been able to interpret law and doctrine under the putative guidance of the *imām*, who will return toward the end of time to fill the world with truth and justice.

Emphasis on the *imām*

On the basis of their doctrine of imamology, the Shīʿah emphasize their idealism and transcendentalism in conscious contrast with Sunnī pragmatism. Thus, whereas the Sunnīs believe in the *ijmāʿ* ("consensus") of the community as the source of decision making and workable knowledge, the Shīʿah believe that knowledge derived from fallible sources is useless and that sure and true knowledge can come only through a contact with the infallible *imām*. Again, in marked contrast to Sunnism, Shīʿism adopted the Muʿtazilite doctrine of the freedom of the human will and the capacity of human reason to know good and evil, although its position on the question of the relationship of faith to works is the same as that of the Sunnīs.

Parallel to the doctrine of an esoteric knowledge, Shīʿism, because of its early defeats and persecutions, also adopted the principle of *taqīyah*, or dissimulation of faith in a hostile environment. Introduced first as a practical principle, *taqīyah*, which is also attributed to ʿAlī and other *imāms*, became an important part of the Shīʿah religious teaching and practice. In the sphere of law, Shīʿism differs from Sunnī law mainly in allowing a temporary marriage, called *mutʿah*, which can be legally contracted for a fixed period of time on the stipulation of a fixed dower.

From a spiritual point of view, perhaps the greatest difference between Shīʿism and Sunnism is the former's introduction into Islām of the passion motive, which is conspicuously absent from Sunnī Islām. The violent death (in 680) of ʿAlī's son, Ḥusayn, at the hands of the Umayyad troops is celebrated with moving orations, passion plays, and processions in which the participants, in a state of emotional frenzy, beat their breasts with heavy chains and sharp instruments, inflicting wounds on their bodies. This passion motive has also influenced the Sunnī masses in Afghanistan and the Indian subcontinent, who participate in passion plays called *taʿziyahs*. Such celebrations are, however, absent from Egypt and North Africa.

Introduction of the passion motive

Although the Shīʿah number only about 40,000,000 (Shīʿism has been the official religion in Iran since the 16th century), Shīʿism has exerted a great influence on Sunnī Islām in several ways. The veneration in which all Muslims hold ʿAlī and his family and the respect shown to ʿAlī's descendants (who are called *sayyids* in the East and *sharīfs* in North Africa) are obvious evidence of this influence.

Ismāʿīlīs. Besides the main body of Twelver (Ithnā ʿAsharīyah) Shīʿah, Shīʿism has produced a variety of more or less extremist sects, the most important of them being the Ismāʿīlī. Instead of recognizing Mūsā as the seventh *imām*, as did the main body of the Shīʿah, the Ismāʿīlīs upheld the

Muslims carrying the *ta'ziyah* to their cremation during a procession commemorating the martyrdom of Ḥusayn, in Jaipur, India.
Foto Features

claims of his elder brother Ismāʿīl. One group of Ismāʿīlīs, called Seveners (Sabʿīyah), considered Ismāʿīl the seventh and last of the *imāms*. The majority of Ismāʿīlīs, however, believed that the imamate continued in the line of Ismāʿīl's descendants. The Ismāʿīlī teaching spread during the 9th century from North Africa to Sind, in India, and the Ismāʿīlī Fāṭimid dynasty succeeded in establishing a prosperous empire in Egypt. Ismāʿīlīs are subdivided into two groups—the Nizārīs, headed by the Aga Khan, and the Mustaʿlīs in Bombay, with their own spiritual head. The Ismāʿīlīs are to be found mainly in East Africa, Pakistan, India, and Yemen.

In their theology, the Ismāʿīlīs have absorbed the most extreme elements and heterodox ideas. The universe is viewed as a cyclic process, and the unfolding of each cycle is marked by the advent of seven "speakers"—messengers of God with Scriptures—each of whom is succeeded by seven "silents"—messengers without revealed scriptures; the last speaker (the Prophet Muḥammad) is followed by seven *imāms* who interpret the Will of God to man and are, in a sense, higher than the Prophet because they draw their knowledge directly from God and not from the Angel of Revelation. During the 10th century, certain Ismāʿīlī intellectuals formed a secret society called the Brethren of Purity, which issued a philosophical encyclopaedia, *The Epistles of the Brethren of Purity*, aiming at the liquidation of positive religions in favour of a universalist spirituality.

The late Aga Khan III (1887–1957) had taken several measures to bring his followers closer to the main body of the Muslims. The Ismāʿīlīs, however, still have not mosques but *jamāʿat khānahs* ("gathering houses"), and their mode of worship bears little resemblance to that of the Muslims generally.

Related sects. Several other sects arose out of the general Shīʿite movement—*e.g.*, the Nuṣayrīs, the Yazīdīs, and the Druzes—which are sometimes considered as independent from Islām. The Druzes arose in the 11th century out of a cult of deification of the Fāṭimid caliph al-Ḥākim.

Druzes and Bahāʾīs

During a 19th-century anticlerical movement in Iran, a certain ʿAlī Moḥammad of Shīrāz appeared, declaring himself to be the Bāb ("Gate"; *i.e.*, to God). At that time the climate in Iran was generally favourable to messianic ideas. He was, however, bitterly opposed by the Shīʿah *ʿulamāʾ* (council of learned men) and was executed in 1850. After his death, his two disciples, Ṣobḥ-e Azal and Bahāʾ Ullāh, broke and went in different directions. Bahāʾ Ullāh eventually declared his religion—stressing a humanitarian pacificism and universalism—to be an independent religion outside Islām. The Bahāʾī faith won a considerable number of converts in North America during the early 20th century (see also in the *Micropædia:* DRUZE and BAHĀʾĪ FAITH).

The Ṣūfīs. Islāmic mysticism, or Ṣūfism, emerged out of early ascetic reactions on the part of certain religiously sensitive personalities against the general worldliness that had overtaken the Muslim community and the purely "externalist" expressions of Islām in law and theology. These persons stressed the Muslim qualities of moral motivation, contrition against overworldliness, and "the state of the heart" as opposed to the legalist formulations of Islām. For a complete exposition of Ṣūfī history, beliefs, and practices, see below *Islāmic mysticism, Ṣūfism.*

OTHER GROUPS

The Aḥmadīyah. In the latter half of the 19th century in Punjab, India, Mirza Ghulam Ahmad claimed to be an inspired prophet. At first a defender of Islām against Christian missionaries, he then later adopted certain doctrines of the Indian Muslim modernist Sayyid Ahmad Khan—namely, that Jesus died a natural death and was not assumed into heaven as the Islāmic orthodoxy believed and that *jihād* "by the sword" had been abrogated and replaced with *jihād* "of the pen." His aim appears to have been to synthesize all religions under Islām, for he declared himself to be not only the manifestation of the Prophet Muḥammad but also the Second Advent of Jesus, as well as Krishna for the Hindus, among other claims. He did not announce, however, any new revelation or new law.

In 1914 a schism over succession occurred among the Aḥmadīyah. One group that seceded from the main body, which was headed by a son of the founder, disowned the prophetic claims of Ghulam Ahmad and established its centre in Lahore (in modern Pakistan). The main body of the Aḥmadīyah (known as the Qadiani, after the village of Qadian, birthplace of the founder and the group's first centre) evolved a separatist organization and, after the partition of India in 1947, moved their headquarters to Rabwah in what was then West Pakistan.

Both groups are noted for their missionary work, particularly in the West and in Africa. Within the Muslim countries, however, there is fierce opposition to the main group because of its claim that Ghulam Ahmad was a prophet (most Muslim sects believe in the finality of prophethood with Muḥammad) and because of its separatist organization. Restrictions were imposed on the Aḥmadīyah in 1974 and again in 1984 by the Pakistani government, which declared that the group was not Muslim and prohibited them from engaging in various Islāmic activities.

The "Black Muslims." After World War II an Islāmic movement arose among blacks in the United States; members called themselves the Nation of Islam, but they were popularly known as Black Muslims. Although they adopted some Islāmic social practices, the group was in large part a black separatist and social protest movement. Their leader, Elijah Muhammad, who claimed to be an inspired prophet, interpreted the doctrine of Resurrection in an unorthodox sense as the revival of oppressed ("dead") peoples. The popular leader and spokesman Malcolm X (el-Hajj Malik el-Shabazz) broke with Elijah Muhammad and adopted more orthodox Islāmic views. He was assassinated in 1965. After the death of Elijah Muhammad in 1975, the group was renamed World Community of Islam in the West and officially abandoned its separatist aims. The name was again changed in the late 1970s, to American Muslim Mission. (F.R./Ed.)

Islāmic mysticism, Ṣūfism

Mysticism is that aspect of Islāmic belief and practice in which Muslims seek to find the truth of divine love and knowledge through direct personal experience of God. It consists of a variety of mystical paths that are designed to ascertain the nature of man and God and to facilitate the experience of the presence of divine love and wisdom in the world.

Islāmic mysticism is called *taṣawwuf* (literally, "to dress in wool") in Arabic, but it has been called Ṣūfism in Western languages since the early 19th century. An abstract word, Ṣūfism derives from the Arabic term for a mystic, *ṣūfī*, which is in turn derived from *ṣūf*, "wool," plausibly a reference to the woollen garment of early Islāmic ascetics. The Ṣūfīs are also generally known as "the poor," *fuqarā*, plural of the Arabic *faqīr*, in Persian *darvīsh*, whence the English words fakir and dervish.

Origins and influence

Though the roots of Islāmic mysticism formerly were supposed to have stemmed from various non-Islāmic sources in ancient Europe and even India, it now seems established that the movement grew out of early Islāmic asceticism that developed as a counterweight to the increasing worldliness of the expanding Muslim community; only later were foreign elements that were compatible with mystical theology and practices adopted and made to conform to Islām.

By educating the masses and deepening the spiritual concerns of the Muslims, Ṣūfism has played an important role in the formation of Muslim society. Opposed to the dry casuistry of the lawyer-divines, the mystics nevertheless scrupulously observed the commands of the divine law. The Ṣūfīs have been further responsible for a large-scale missionary activity all over the world, which still continues. Ṣūfīs have elaborated the image of the prophet Muḥammad—the founder of Islām—and have thus largely influenced Muslim piety by their Muḥammad-mysticism. Without the Ṣūfī vocabulary, Persian and other literatures related to it, such as Turkish, Urdu, Sindhi, Pashto, and Panjabi, would lack their special charms. Through the poetry of these literatures mystical ideas spread widely among the Muslims. In some countries Ṣūfī leaders were also active politically.

HISTORY

Islāmic mysticism had several stages of growth, including (1) the appearance of early asceticism, (2) the development of a classical mysticism of divine love, and (3) the rise and proliferation of fraternal orders of mystics. Despite these general stages, however, the history of Islāmic mysticism is largely a history of individual mystic experience.

The first stage of Ṣūfism appeared in pious circles as a reaction against the worldliness of the early Umayyad period (AD 661–749). From their practice of constantly meditating on the Qur'ānic words about Doomsday, the ascetics became known as "those who always weep" and those who considered this world "a hut of sorrows." They were distinguished by their scrupulous fulfillment of the injunctions of the Qur'ān and tradition, by many acts of piety, and especially by a predilection for night prayers.

Classical mysticism. The introduction of the element of love, which changed asceticism into mysticism, is ascribed to Rābi'ah al-'Adawīyah (died 801), a woman from Basra who first formulated the Ṣūfī ideal of a love of God that was disinterested, without hope for paradise and without fear of hell. In the decades after Rābi'ah, mystical trends grew everywhere in the Islāmic world, partly through an exchange of ideas with Christian hermits. A number of mystics in the early generations had concentrated their efforts upon *tawakkul*, absolute trust in God, which became a central concept of Ṣūfism. An Iraqi school of mysticism became noted for its strict self-control and psychological insight. The Iraqi school was initiated by al-Muḥāsibī (died 857)—who believed that purging the soul in preparation for companionship with God was the only value of asceticism. Its teachings of classical sobriety and wisdom were perfected by Junayd of Baghdad (died 910), to whom all later chains of the transmission of doctrine and legitimacy go back. In an Egyptian school of Ṣūfism, the Nubian Dhū an-Nūn (died 859) reputedly introduced the technical term *ma'rifah* ("interior knowledge"), as contrasted to learnedness; in his hymnical prayers he joined all nature in the praise of God—an idea based on the Qur'ān and later elaborated in Persian and Turkish poetry. In the Iranian school, Abū Yazīd al-Bisṭāmī (died 874) is usually considered to have been representative of the important doctrine of annihilation of the self, *fanā'* (see below); the strange symbolism of his sayings prefigures part of the terminology of later mystical poets. At the same time the concept of divine love became more central, especially among the Iraqi Ṣūfīs. Its main representatives are Nūrī, who offered his life for his brethren, and Sumnūn "the Lover."

Mystical love

The first of the theosophical speculations based on mystical insights about the nature of man and the essence of the Prophet were produced by such Ṣūfīs as Sahl at-Tustarī (died *c.* 896). Some Hellenistic ideas were later adopted by al-Ḥakīm at-Tirmidhī (died 898). Sahl was the master of al-Ḥusayn ibn Manṣūr al-Ḥallāj, who has become famous for his phrase *anā al-ḥaqq*, "I am the Creative Truth" (often rendered "I am God"), which was later interpreted in a pantheistic sense but is, in fact, only a condensation of his theory of *huwa huwa* ("He he"): God loved himself in his essence, and created Adam "in his image." Ḥallāj was executed in 922 in Baghdad as a result of his teachings; he is, for later mystics and poets, the "martyr of Love" par excellence, the enthusiast killed by the theologians. His few poems are of exquisite beauty; his prose, which contains an outspoken Muḥammad-mysticism—*i.e.*, mysticism centred on the prophet Muḥammad—is as beautiful as it is difficult.

Ṣūfī thought was in these early centuries transmitted in small circles. Some of the *shaykh*s, Ṣūfī mystical leaders or guides of such circles, were also artisans. In the 10th century, it was deemed necessary to write handbooks about the tenets of Ṣūfism in order to soothe the growing suspicions of the orthodox; the compendiums composed in Arabic by Abū Ṭālib Makkī, Sarrāj, and Kalābādhī in the late 10th century, and by Qushayrī and, in Persian, by Hujvīrī in the 11th century reveal how these authors tried to defend Ṣūfism and to prove its orthodox character. It should be noted that the mystics belonged to all schools of Islāmic law and theology of the times.

The last great figure in the line of classical Ṣūfism is Abū Ḥāmid al-Ghazālī (died 1111), who wrote, among numerous other works, the *Iḥyā' 'ulūm ad-dīn* ("The Revival of the Religious Sciences"), a comprehensive work that established moderate mysticism against the growing theosophical trends—which tended to equate God and the world—and thus shaped the thought of millions of Muslims. His younger brother, Aḥmad al-Ghazālī, wrote one of the subtlest treatises (*Sawāniḥ*; "Occurrences" [*i.e.*, stray thoughts]) on mystical love, a subject that then became the main subject of Persian poetry.

Rise of fraternal orders. Slightly later, mystical orders (fraternal groups centring around the teachings of a leader-founder) began to crystallize. The 13th century, though politically overshadowed by the invasion of the Mongols into the Eastern lands of Islām and the end of the 'Abbāsid caliphate, was also the golden age of Ṣūfism: the Spanish-born Ibn al'Arabī created a comprehensive theosophical system (concerning the relation of God and the world) that was to become the cornerstone for a theory of "Unity of Being." According to this theory all existence is one, a manifestation of the underlying divine reality. His Egyptian contemporary Ibn al-Fāriḍ wrote the finest mystical poems in Arabic. Two other important mystics, who died *c.* AD 1220, were a Persian poet, Farīd od-Dīn 'Aṭṭar, one of the most fertile writers on mystical topics, and a Central Asian master, Najmuddīn Kubrā, who presented elaborate discussions of the psychological experiences through which the mystic adept has to pass.

The greatest mystical poet in the Persian language, Jalāl ad-Dīn ar-Rūmī (1207–73), was moved by mystical love to compose his lyrical poetry that he attributed to his mystical beloved, Shams ad-Dīn of Tabriz, as a symbol of their union. Rūmī's didactic poem *Maṡnavī* in about 26,000 couplets—a work that is for the Persian-reading mystics second in importance only to the Qur'ān—is an encyclopaedia of mystical thought in which everyone can find his own religious ideas. Rūmī inspired the organization of the whirling dervishes—who sought ecstasy through an elaborate dancing ritual, accompanied by superb music. His younger contemporary Yunus Emre inaugurated Turkish mystical poetry with his charming verses that were transmitted by the Bektāshīyah (Bektaşi) order of dervishes and are still admired in modern Turkey. In Egypt, among many other mystical trends, an order—known as Shādhilīyah—was founded by ash-Shādhilī (died 1258); its main literary representative, Ibn 'Aṭā' Allāh of Alexandria, wrote sober aphorisms (*ḥikam*).

The mystical poetry of Jalāl ad-Dīn ar-Rūmī

At that time, the basic ideals of Ṣūfism permeated the whole world of Islām; and at its borders as, for example, in

India, Ṣūfīs largely contributed to shaping Islāmic society. Later some of the Ṣūfīs in India were brought closer to Hindu mysticism by an overemphasis on the idea of divine unity which became almost monism—a religiophilosophic perspective according to which there is only one basic reality, and the distinction between God and the world (and man) tends to disappear. The syncretistic attempts of the Mughal emperor Akbar (died 1605) to combine different forms of belief and practice, and the religious discussions of the crown prince Dārā Shukōh (executed for heresy, 1659) were objectionable to the orthodox. Typically, the countermovement was again undertaken by a mystical order, the Naqshbandīyah, a Central Asian fraternity founded in the 14th century. Contrary to the monistic trends of the school of *waḥdat al-wujūd* ("existential unity of being"), the later Naqshbandīyah defended the *waḥdat ash-shuhūd* ("unity of vision"), a subjective experience of unity, occurring only in the mind of the believer, and not as an objective experience. Aḥmad Sirhindī (died 1624) was the major protagonist of this movement in India. His claims of sanctity were surprisingly daring: he considered himself the divinely invested master of the universe. His refusal to concede the possibility of union between man and God (characterized as "servant" and "Lord") and his sober law-bound attitude gained him and his followers many disciples, even at the Mughal court and as far away as Turkey. In the 18th century, Shāh Walī Allāh of Delhi was connected with an attempt to reach a compromise between the two inimical schools of mysticism; he was also politically active and translated the Qur'ān into Persian, the official language of Mughal India. Other Indian mystics of the 18th century, such as Mīr Dard, played a decisive role in forming the newly developing Urdu poetry.

Trends in modern Ṣūfism

In the Arabic parts of the Islāmic world, only a few interesting mystical authors are found after 1500. They include ash-Shaʿrānī in Egypt (died 1565) and the prolific writer ʿAbd al-Ghanī an-Nābulusī in Syria (died 1731). Turkey produced some fine mystical poets in the 17th and 18th centuries. The influence of the mystical orders did not recede; rather new orders came into existence, and most literature was still tinged with mystical ideas and expressions. Political and social reformers in the Islāmic countries have often objected to Ṣūfism because they have generally considered it as backward, hampering the free development of society. Thus, the orders and dervish lodges in Turkey were closed by Kemal Atatürk in 1925. Yet, their political influence is still palpable, though under the surface. Such modern Islāmic thinkers as the Indian philosopher Muḥammad Iqbāl have attacked traditional monist mysticism and have gone back to the classical ideals or divine love as expressed by Ḥallāj and his contemporaries. The activities of modern Muslim mystics in the cities are mostly restricted to spiritual education.

SUFI LITERATURE

Though a prophetic saying (Ḥadīth) claims that "he who knows God becomes silent," the Ṣūfīs have produced a literature of impressive extent and could defend their writing activities with another Ḥadīth: "He who knows God talks much." The first systematic books explaining the tenets of Ṣūfism date from the 10th century; but earlier, Muḥāsibī had already written about spiritual education, Ḥallāj had composed meditations in highly concentrated language, and many Ṣūfīs had used poetry for conveying their experiences of the ineffable mystery or had instructed their disciples in letters of cryptographic density. The accounts of Ṣūfism by Sarrāj and his followers, as well as the *ṭabaqāt* (biographical works) by Sulamī, Abū Nu'aym al-Iṣfahānī, and others, together with some biographies of individual masters, are the sources for knowledge of early Ṣūfism.

Early mystical commentaries on the Qur'ān are only partly extant, often preserved in fragmentary quotations in later sources. With the formation of mystical orders, books about the behaviour of the Ṣūfī in various situations became important, although this topic had already been touched on in such classical works as *Adāb al-murīdīn* ("The Adepts' Etiquette") by Abū Najīb as-Suhrawardī (died 1168), the founder of the Suhrawardīyah order and uncle of the author of the oft translated *'Awārif al-ma'ārif* ("The Well-known Sorts of Knowledge"). The theosophists had to condense their systems in readable form; Ibn al-'Arabī's *al-Futūḥāt al-Makkīyah* ("The Meccan Revelations") is the textbook of *waḥdat al-wujūd* (God and creation as two aspects of one reality); his smaller work on the peculiar character of the prophets—*Fuṣūṣ al-ḥikam* ("The Bezels—or cutting edges—of Wisdom")—became even more popular.

Later mystics commented extensively upon the classical sources and, sometimes, translated them into their mother tongues. A literary type that has flourished especially in India since the 13th century is the *malfūẓāt,* a collection of sayings of the mystical leader, which are psychologically interesting and allow glimpses into the political and social situation of the Muslim community. Collections of letters of the *shaykh*s are similarly revealing. Ṣūfī literature abounds in hagiography, either biographies of all known saints from the Prophet to the day of the author, or of saints of a specific order, or of those who lived in a certain town or province, so that much information on the development of Ṣūfī thought and practice is available if sources are critically sifted.

Poetical, national, and regional literature

The greatest contribution of Ṣūfism to Islāmic literature, however, is poetry—beginning with charming, short Arabic love poems (sometimes sung for a mystical concert, *samā'*) that express the yearning of the soul for union with the beloved. The love-relation prevailing in most Persian poetry is that between a man and a beautiful youth; less often, as in the writings of Ibn al-'Arabī and Ibn al-Fāriḍ, eternal beauty is symbolized through female beauty; in Indo-Muslim popular mystical songs the soul is the loving wife, God the longed-for husband. Long mystic–didactic poems (*masnavīs*) were written to introduce the reader to the problems of unity and love by means of allegories and parables. After Sanā'ī's (died 1131?) *Ḥadīqat al-ḥaqīqah wa sharī'at aṭ-ṭariqah* ("The Garden of Truth and the Law of Practice"),came 'Aṭṭar's *Manṭeq oṭ-ṭeyr* ("The Birds' Conversation") and Rūmī's *Mas̄navī-ye ma'navī* ("Spiritual Couplets"). These three works are the sources that have furnished poets for centuries with mystical ideas and images. Typical of Ṣūfī poetry is the hymn in praise of God, expressed in chains of repetitions.

The mystics also contributed largely to the development of national and regional literatures, for they had to convey their message to the masses in their own languages: in Turkey as well as in the Panjabi-, the Sindhi-, and the Urdu-speaking areas of South Asia, the first true religious poetry was written by Ṣūfīs, who blended classical Islāmic motifs with inherited popular legends and used popular rather than Persian metres. Ṣūfī poetry expressing divine love and mystical union through the metaphors of profane love and union often resembled ordinary worldly love poetry; and nonmystical poetry made use of the Ṣūfī vocabulary, thus producing an ambiguity that is felt to be one of the most attractive and characteristic features of Persian, Turkish, and Urdu literatures. Ṣūfī ideas thus permeated the hearts of all those who hearkened to poetry. An example is al-Ḥusayn ibn Manṣūr al-Ḥallāj, the 10th-century martyr–mystic, who is as popular in modern progressive Urdu poetry as he was with the "God-intoxicated" Ṣūfīs; he has been converted into a symbol of suffering for one's ideals.

SUFI THOUGHT AND PRACTICE

Important aspects. The mystics drew their vocabulary largely from the Qur'ān, which for Muslims contains all divine wisdom and has to be interpreted with ever-increasing insight. In the Qur'ān, mystics found the threat of the Last Judgment, but they also found the statement that God "loves them and they love him," which became the basis for love-mysticism. Strict obedience to the religious law and imitation of the Prophet were basic for the mystics. By rigid introspection and mental struggle the mystic tried to purify his baser self from even the smallest signs of selfishness, thus attaining *ikhlāṣ,* absolute purity of intention and act. *Tawakkul* (trust in God) was sometimes practiced to such an extent that every thought of tomorrow was considered irreligious. "Little sleep, little talk, little food" were fundamental; fasting became one of the most important preparations for the spiritual life.

Central concern of Ṣūfism

The central concern of the Ṣūfis, as of every Muslim, was *tawḥīd*, the witness that "There is no deity but God." This truth had to be realized in the existence of each individual, and so the expressions differ: early Ṣūfism postulated the approach to God through love and voluntary suffering until a unity of will was reached; Junayd spoke of "recognizing God as He was before creation"; God is seen as the One and only actor; He alone "has the right to say 'I'." Later, *tawḥīd* came to mean the knowledge that there is nothing existent but God, or the ability to see God and creation as two aspects of one reality, reflecting each other and depending upon each other (*waḥdat al-wujūd*).

The mystics realized that beyond the knowledge of outward sciences intuitive knowledge was required in order to receive that illumination to which reason has no access. *Dhawq,* direct "tasting" of experience, was essential for them. But the inspirations and "unveilings" that God grants such mystics by special grace must never contradict the Qur'ān and tradition and are valid only for the person concerned. Even the Malāmatīs, who attracted public contempt upon themselves by outwardly acting against the law, in private life strictly followed the divine commands. Mystics who expressed in their poetry their disinterest in, and even contempt of, the traditional formal religions never forgot that Islām is the highest manifestation of divine wisdom.

The idea of the manifestation of divine wisdom was also connected with the person of the prophet Muḥammad. Though early Ṣūfism had concentrated upon the relation between God and the soul, from AD 900 onward a strong Muḥammad-mysticism developed. In the very early years, the alleged divine address to the Prophet—"If thou hadst not been I had not created the worlds"—was common among Ṣūfis. Muḥammad was said to be "Prophet when Adam was still between water and clay." Muḥammad is also described as light from light, and from his light all the prophets are created, constituting the different aspects of this light. In its fullness such light radiated from the historical Muḥammad and is partaken of by his posterity and by the saints; for Muḥammad has the aspect of sanctity in addition to that of prophecy. An apocryphal tradition makes even God attest: "I am Aḥmad (= Muḥammad) without 'm' (*i.e.,* Aḥad, 'One')."

The *walī*, or saint

A mystic may also be known as *walī*. By derivation the word *walī* ("saint") means "one in close relation; friend." The *awlīyā'* (plural of *walī*) are "friends of God who have no fear nor are they sad." Later the term *walī* came to denote the Muslim mystics who had reached a certain stage of proximity to God, or those who had reached the highest mystical stages. They have their "seal" (*i.e.,* the last and most perfect personality in the historical process; with this person, the evolution has found its end—as in Muḥammad's case), just as the prophets have. Woman saints are found all over the Islāmic world.

The invisible hierarchy of saints consists of the 40 *abdāl* ("substitutes"; for when any of them dies another is elected by God from the rank and file of the saints), seven *awtād* ("stakes," or "props," of faith), three *nuqabā'* ("leader"; "one who introduces people to his master"), headed by the *quṭb* ("axis, pole"), or *ghawth* ("help")—titles claimed by many Ṣūfī leaders. Saint worship is contrary to Islām, which does not admit of any mediating role for human beings between man and God; but the cult of living and even more of dead saints—visiting their tombs to take vows there—responded to the feeling of the masses, and thus a number of pre-Islāmic customs were absorbed into Islām under the cover of mysticism. The advanced mystic was often granted the capacity of working miracles called *karāmāt* (*charismata* or "graces"); not *mu'jizāt* ("that which men are unable to imitate"), like the miracles of the prophets. Among them are "cardiognosia" (knowledge of the heart), providing food from the unseen, presence in two places at the same time, and help for the disciples, be they near or far. In short, a saint is one "whose prayers are heard" and who has *taṣarruf,* the power of materializing in this world possibilities that still rest in the spiritual world. Many great saints, however, considered miracle working as a dangerous trap on the path that might distract the Ṣūfī from his real goal.

The path. The path (*ṭarīqah*) begins with repentance. A mystical guide (*shaykh, pīr*) accepts the seeker as disciple (*murīd*), orders him to follow strict ascetic practices, and suggests certain formulas for meditation. It is said that the disciple should be in the hands of the master "like a corpse in the hand of the washer." The master teaches him constant struggle (the real "Holy War") against the lower soul, often represented as a black dog, which should, however, not be killed but merely tamed and used in the way of God. The mystic dwells in a number of spiritual stations (*maqām*), which are described in varying sequence, and, after the initial repentance, comprise abstinence, renunciation, and poverty—according to Muḥammad's saying, "Poverty is my pride"; poverty was sometimes interpreted as having no interest in anything apart from God, the Rich One, but the concrete meaning of poverty prevailed, which is why the mystic is often denoted as "poor," fakir or dervish. Patience and gratitude belong to higher stations of the path, and consent is the loving acceptance of every affliction.

On his way to illumination the mystic will undergo such changing spiritual states (*ḥāl*) as *qabḍ* and *basṭ,* constraint and happy spiritual expansion, fear and hope, and longing and intimacy, which are granted by God and last for longer or shorter periods of time, changing in intensity according to the station in which the mystic is abiding at the moment. The way culminates in *ma'rifah* ("interior knowledge," "gnosis") or in *maḥabbah* ("love"), the central subject of Ṣūfism since the 9th century, which implies a union of lover and beloved, and was therefore violently rejected by the orthodox, for whom "love of God" meant simply obedience. The final goal is *fanā'* ("annihilation"), primarily an ethical concept of annihilating one's own qualities, according to the prophetic saying "Take over the qualities of God," but slowly developing into a complete extinction of the personality. Some mystics taught that behind this negative unity where the self is completely effaced, the *baqā',* ("duration, life in God") is found: the ecstatic experience, called intoxication, is followed by the "second sobriety"; *i.e.,* the return of the completely transformed mystic into this world where he acts as a living witness of God or continues the "journey in God." The mystic has reached *ḥaqīqah* ("realty"), after finishing the *ṭarīqah* ("path"), which is built upon the *sharī'ah* ("law"). Later, the disciple is led through *fanā' fī ashshaykh* ("annihilation in the master") to *fanā' fiar-Rasūl* ("annihilation in the Prophet") before reaching, if at all, *fanā' fi-Allāh* ("annihilation in God").

The practice of *dhikr*

One of the means used on the path is the ritual prayer, or *dhikr* ("remembrance"), derived from the Qur'ānic injunction "And remember God often" (*sūrah* 62:10). It consists in a repetition of either one or all of the most beautiful names of God, of the name "Allāh," or of a certain religious formula, such as the profession of faith: "There is no God but Allāh and Muḥammad is his prophet." The rosary with 99 or 33 beads was in use as early as the 8th century for counting the thousands of repetitions. Man's whole being should eventually be transformed into remembrance of God.

In the mid-9th century some mystics introduced sessions with music and poetry recitals (*samā'*) in Baghdad in order to reach the ecstatic experience—and since then debates about the permissibility of *samā',* filling many books, have been written. Narcotics were used in periods of degeneration, coffee by the "sober" mystics (first by the Shādhilīyah after 1300).

Besides the wayfarers (*sālik*) on the path, Ṣūfīs who have no master but are attracted solely by divine grace are also found; they are called Uwaysī, after Uways al-Qaranī, the Yemenite contemporary of the Prophet who never saw him but firmly believed in him. There are also the so-called *majdhūb* ("attracted") who are often persons generally agreed to be more or less mentally deranged.

Symbolism in Ṣūfism. The divine truth was at times revealed to the mystic in visions, auditions, and dreams, in colours and sounds, but to convey these nonrational and ineffable experiences to others the mystic had to rely upon such terminology of worldly experience as that of love and intoxication—often objectionable from the orthodox viewpoint. The symbolism of wine, cup, and cupbearer, first expressed by Abū Yazīd al-Bisṭāmī in the 9th century, became popular everywhere, whether in the verses of the Arab Ibn al-Fāriḍ, or the Persian 'Irāqī, or the Turk Yunus Emre, and their followers. The hope for the union of the soul with the divine had to be expressed through images of human yearning and love. The love for lovely boys in which the divine beauty manifests itself—according to the alleged Ḥadīth "I saw my Lord in the shape of a youth with a cap awry"—was commonplace in Persian poetry. Union was described as the submersion of the drop in the ocean, the state of the iron in the fire, the vision of penetrating light, or the burning of the moth in the candle (first used by Ḥallāj). Worldly phenomena were seen as black tresses veiling the radiant beauty of the divine countenance. The mystery of unity and diversity was symbolized, for example, under the image of mirrors that reflect the different aspects of the divine, or as prisms colouring the pure light. Every aspect of nature was seen in relation to God. The symbol of the soulbird—in which the human soul is likened to a flying bird—known everywhere, was the centre of 'Aṭṭar's *Manṭeq ot-ṭeyr* ("The Birds' Conversation"). The predilection of the mystical poets for the symbolism of the nightingale and rose (the red rose = God's perfect beauty; nightingale = soul; first used by Baqli [died 1206]) stems from the soulbird symbolism. For spiritual education, symbols taken from medicine (healing of the sick soul) and alchemy (changing of base matter into gold) were also used. Many descriptions that were originally applied to God as the goal of love were, in later times, used also for the Prophet, who is said to be like the "dawn between the darkness of the material world and the sun of Reality."

Symbolism of union with the divine

Allusions to the Qur'ān were frequent, especially so to verses that seem to imply divine immanence (God's presence in the world), such as "Whithersoever ye turn, there is the Face of God" (*sūrah* 2:109), or that God is "Closer than your neck-vein" (*sūrah* 50:8). *Sūrah* 7:172—*i.e.*, God's address to the uncreated children of Adam ("Am I not your Lord" [*alastu birabbikum*])—came to denote the pre-eternal love relation between God and man. As for the prophets before Muḥammad, the vision of Moses was considered still imperfect, for the mystic wants the actual vision of God, not His manifestation through a burning bush. Abraham, for whom fire turned into a rose garden, resembles the mystic in his afflictions; Joseph, in his perfect beauty, the mystical beloved after whom the mystic searches. The apocryphal traditions used by the mystics are numerous; such as "Heaven and earth do not contain me, but the heart of my faithful servant contains Me"; and the possibility of a relation between man and God is also explained by the traditional idea: "He (God) created Adam in His image."

THEOSOPHICAL SUFISM

Ṣūfism, in its beginnings a practical method of spiritual education and self-realization, grew slowly into a theosophical system by adopting traditions of Neoplatonism, the Hellenistic world, Gnosticism (an ancient esoteric religiophilosophical movement that viewed matter as evil and spirit as good), and spiritual currents from Iran and various countries in the ancient agricultural lands from the eastern Mediterranean to Iraq. One master who contributed to this development was the Persian as-Suhrawardī, called al-Maqtūl ("killed"), executed in 1191 in Aleppo. To him is attributed the philosophy of *ishrāq* ("illumination"), and he claimed to unite the Persian (Zoroastrian) and Egyptian (Hermetic) traditions. His didactic and doctrinal works in Arabic among other things taught a complicated angelology (theory of angels); some of his smaller Persian treatises depict the journey of the soul across the cosmos; the "Orient" (East) is the world of pure lights and archangels, the "Occident" (West) that of darkness and matter; and man lives in the "Western exile."

At the time of Suhrawardī's death the greatest representative of theosophic Ṣūfism was in his 20s: Ibn al-'Arabī, born at Murcia, Spain, where speculative tendencies had been visible since Ibn Masarrah's philosophy (died 931). Ibn al-'Arabī was instructed in mysticism by two Spanish woman saints. Performing the traditional pilgrimage to Mecca, he met there an accomplished young Persian lady who represented for him the divine wisdom. This experience resulted in the charming poems of the *Tarjumān al-ashwāq* ("Interpreter of Yearning"), which the author later explained mystically. Ibn al-'Arabī composed at least 150 volumes. His magnum opus is *al-Futūḥāt al-Makkīyah* ("The Meccan Revelations") in 560 chapters, in which he expounds his theory of unity of being.

Contributions of Ibn al-'Arabī

The substance of theosophic Ṣūfism is as follows. According to the Ḥadīth *qudsī*, or "holy tradition"—"I was a hidden treasure and wanted to be known"—the absolute, or God, yearned in his loneliness for manifestation and created the world by effusing being upon the heavenly archetypes, a "theophany (a physical manifestation of deity) through God's imaginative power." The universe is annihilated and created every moment. Every divine name is reflected in a named one. The world and God are said to be like ice and water, or like two mirrors contemplating themselves in each other, joined by a sympathetic union. The Prophet Muḥammad is the universal man, the perfect man, the total theophany of the divine names, the prototype of creation. Muḥammad is the "word," each particular dimension of which is identified with a prophet, and he is also the model for the spiritual realization of the possibilities of man. The mystic has to pass the stages of the Qur'ānic prophets as they are explained in the *Fuṣūṣ al-ḥikam* ("Bezels of Wisdom") until he becomes united with the *ḥaqīqa Muḥammadīya* (the first individualization of the divine in the "Muḥammadan Reality"). Man can have vision only of the form of the faith he professes, and Ibn al-'Arabī's oft-quoted verse, "I follow the religion of love wherever its camels turn," with its seeming religious tolerance means, as S.H. Nasr puts it: "the form of God is for him no longer the form of this or that faith exclusive of all others but his own eternal form which he encounters." The theories of the perfect man were elaborated by Jīlī (died *c.* 1424) in his compendium *Al-insān al-kāmil* ("The Perfect Man") and became common throughout the Muslim world.

Ibn al-'Arabī's theosophy has been attacked by orthodox Muslims and mystics of the "sober" school as incongruent with Islām because "a thoroughly monistic system cannot take seriously the objective validity of moral standards." Even the adversaries of the "greatest master" could not, however, help using part of his terminology. Innumerable mystics and poets propagated his ideas, though they only partly understood them, and this circumstance led also to a misinterpretation of the data of early Ṣūfism in the light of existential monism. Later Persian poetry is permeated by the pantheistic feeling of *hama ost* ("everything is He").

Ibn al-'Arabī's contemporary in Egypt, the poet Ibn al-Fāriḍ, is usually mentioned together with him; Ibn al-Fāriḍ, however, is not a systematic thinker but a full-fledged poet who used the imagery of classical Arabic poetry to describe the state of the lover in extremely artistic verses and has given, in his *Tā'iyat al-kubrā* ("Poem of the Journey"), glimpses of the way of the mystic, using, as many poets before and after him did, for example, the image of the shadow play for the actions of the creatures who are dependent upon the divine playmaster. His unifying experience is personal and is not the expression of a theosophical system.

SUFI ORDERS

Organization. Mystical life was first restricted to the relation between a master and a few disciples; the foundations of a monastic system were laid by the Persian Abū Sa'īd ebn Abī ol-Kheyr (died 1049), but real orders or fraternities came into existence only from the 12th century onward: 'Abd al-Qādir al-Jīlānī (died 1166) gathered the first and still most important order around himself; then followed the Suhrawardīyah, and the 13th century saw the formation of large numbers of different orders in the East (for example, Kubrawīya in Khvārezm) and West (Shādhilīyah). Thus, Ṣūfism ceased to be the way of the chosen few and influenced the masses. A strict ritual was

Initiation and investiture

elaborated: when the adept had found a master for whom he had to feel a preformed affinity, there was an initiation ceremony in which he swore allegiance (*bay'at*) into the master's hand; similarities to the initiation in Ismā'īlism, the 9th-century sect, and in the guilds suggest a possible interaction. The disciple (*murīd*) had to undergo a stern training; he was often ordered to perform the lowest work in the community, to serve the brethren, to go out to beg (many of the old monasteries subsisted upon alms). A seclusion period of 40 days under hard conditions was common for the adepts in most orders.

Investiture with the *khirqah*, the frock of the master, originally made from shreds and patches, was the decisive act by which the disciple became part of the *silsilah*, the chain of mystical succession and transmission, which leads back—via Junayd—to the Prophet himself and differs in every order. Some mystical leaders claimed to have received their *khirqah* directly from al-Khiḍr, a mysterious immortal saint.

In the earliest times, allegiance was sworn exclusively to one master who had complete power over the disciple, controlling each of his movements, thoughts, visions, and dreams; but later many Ṣūfis got the *khirqah* from two or more *shaykhs*. There is consequently a differentiation between the *shaykh at-tarbiyah*, who introduces the disciple into the ritual, forms, and literature of the order, and the *shaykh aṣ-ṣuḥbah*, who steadily watches him and with whom the disciple lives. Only a few members of the fraternity remained in the centre (*dargāh, khānqāh, tekke*), close to the *shaykh*, but even those were not bound to celibacy. Most of the initiated returned to their daily life and partook in mystic services only during certain periods. The most mature disciple was invested as *khalīfah* ("successor") to the *shaykh* and was often sent abroad to extend the activities of the order. The *dargāhs* were organized differently in the various orders; some relied completely upon alms, keeping their members in utmost poverty; others were rich, and their *shaykh* was not very different from a feudal lord. Relations with rulers varied—some masters refused contacts with the representatives of political power; others did not mind friendly relations with the grandees.

Discipline and ritual. Each order has peculiarities in its ritual. Most start the instruction with breaking the lower soul; others, such as the later Naqshbandīyah, stress the purification of the heart by constant *dhikr* ("remembrance") and by discourse with the master (*ṣuḥbah*). The forms of *dhikr* vary in the orders. Many of them use the word Allāh, or the profession of faith with its rhythmical wording, sometimes accompanied by movements of the body, or by breath control up to complete holding of the breath. The Mawlawīs, the whirling dervishes, are famous for their dancing ritual, an organized variation of the earlier *samā'* practices, which were confined to music and poetry. The Rifā'is, the so-called Howling Dervishes, have become known for their practice of hurting themselves while in an ecstatic state that they reach in performing their loud *dhikr*. (Such practices that might well degenerate into mere jugglery are not approved by most orders.) Some orders also teach the *dhikr khafī*, silent repetition of the formulas, and meditation, concentrating upon certain fixed points of the body; thus the Naqshbandīs do not allow any emotional practices and prefer contemplation to ecstasy, perhaps as a result of Buddhist influence from Central Asia. Other orders have special prayers given to the disciples, such as the protective *ḥizb al-baḥr* ("The protective armour of the sea"; *i.e.*, for seafaring people—then extended to all travellers) in the Shādhilīyah order. Most of them prescribe for their disciples additional prayers and meditation at the end of each ritual prayer.

Function and role in Islāmic society. The orders formed an excellent means of bringing together the spiritually interested members of the community. They acted as a counterweight against the influence of hairsplitting lawyer-divines and gave the masses an emotional outlet in enthusiastic celebrations (*'urs*, "marriage") of the anniversaries of the deaths of founders of mystic orders or similar festivals in which they indulged in music and joy. The orders were adaptable to every social level; thus, some of them were responsible for adapting a number of un-Islāmic folkloristic practices such as veneration of saints.

Adaptation to levels of society

Their way of life often differed so much from Islāmic ideals that one distinguishes in Iran and India between orders *bā shar'* (law-bound) and *bī shar'* (not following the injunctions of the Qur'ān). Some orders were more fitting for the rural population, such as the Aḥmadīyah (after Aḥmad al-Badawī; died 1286) in Egypt. The Aḥmadīyah, however, even attracted some Mamlūk rulers. The Turkish Bektāshīyah (Haci Bektaş, early 14th century), together with strange syncretistic cults, showed a prevalence of the ideals of the Shī'ites (from Shī'ah—the followers of 'Alī, son-in-law of the prophet Muḥammad, whose descendants claimed to be rightful successors to the religious leadership of Islām). The figure of 'Alī played a role also in other fraternities, and the relations between Ṣūfism in the 14th and 15th centuries and the Shī'ah still have to be explored, as is also true of the general influence of Shī'ite ideas on Ṣūfism. Other orders, such as the Shādhilīyah, an offshoot of which still plays an important role among Egyptian officials and employees, are typically middle class. This order demands not a life in solitude but strict adherence to one's profession and fulfillment of one's duty. Still other orders were connected with the ruling classes, such as, for a time, the Chishtīyah in Mughal India, and the Mawlawīyah, whose leader had to invest the Ottoman sultan with the sword. The Mawlawīyah is also largely responsible for the development of classical Turkish poetry, music, and fine arts, just as the Chishtīyah contributed much to the formation of classical Indo-Muslim music.

Missionary activity

The main contribution of the orders, however, is their missionary activity. The members of different orders who settled in India from the early 13th century attracted thousands of Hindus by their example of love of both God and their own brethren and by preaching the equality of men. Missionary activity was often joined with political activity, as in 17th- and 18th-century Central Asia, where the Naqshbandīyah exerted strong political influence. In North Africa the Tijānīyah, founded in 1781, and the Sanūsīyah, active since the early 19th century, both heralded Islām and engaged in politics; the Sanūsīyah fought against Italy, and the former king of Libya was the head of the order. The Tijānīyah extended the borders of Islām toward Senegal and Nigeria, and their representatives founded large kingdoms in West Africa. Their influence, as well as that of the Qādirīyah (see below), is still an important sociopolitical factor in those areas.

Geographical extent of Ṣūfī orders. It would be impossible to number the members of mystical orders in the Islāmic world. Even in such countries as Turkey, where the orders have been banned since 1925, many people still cling to the mystical tradition and feel themselves to be links in the spiritual chains of the orders and try to implement their ideals in modern society. The most widely spread group is, no doubt, the Qādirīyah, whose adherents are found from West Africa to India—the tomb of 'Abd al-Qādir al-Jīlānī in Baghdad still being a place of pilgrimage. The areas where the Sanūsīyah live are restricted to the Maghrib, the Atlas Massif, and the coastal plain from Morocco to Tunisia, whereas the Tijānīyah has some offshoots in Turkey. Such rural orders as the Egyptian Aḥmadīyah and Dasūqīyah (named after Ibrāhīm ad-Dasūqī; died 1277) are bound to their respective countries, as are the Mawlawīs and Bektāshīyah to the realms of the former Ottoman Empire. The Bektāshīyah had gained political importance in the empire because of its relations with the Janissaries, the standing army. Albania, since 1929, has had a strong and officially recognized group of Bektāshīyah who were even granted independent status after World War II. The Shaṭṭārīyah (derived from 'Abd ash-Shaṭṭār; died 1415) extends from India to Java, whereas the Chishtīyah (derived from Khwājah Mu'īnud-Dīn Chishtīp; died 1236 in Ajmer) and Suhrawardīyah remain mainly inside the Indo-Pakistan subcontinent. The Kubrāwīyah reached Kashmir through 'Alī Hama-dhānī (died 1385), a versatile author, but the order later lost its influence.

The great variety of possible forms may be seen by comparing the Haddāwah, vagabonds in Morocco, who "do not spoil God's day by work" and the Shādhilīyah with a sober attitude toward professional life and careful introspection. Out of the Shādhilīyah developed the austere Darqāwīyah, who, in turn, produced the ʿAlāwīyah, whose master has attracted even a number of Europeans. The splitting up and formation of suborders is a normal process, but most of the subgroups have only local importance. The High Ṣūfī Convent in Egypt counts 60 registered orders.

SIGNIFICANCE

Ṣūfism has helped to shape large parts of Muslim society. The orthodox disagree with such aspects of Ṣūfism as saint worship, visiting of tombs, musical performances, miracle mongering, degeneration into jugglery, and the adaptation of pre-Islāmic and un-Islāmic customs; and the reformers object to the influences of the monistic interpretation of Islām upon moral life and human activities. The importance given to the figure of the master is accused of yielding negative results; the *shaykh* as the almost infallible leader of his disciples and admirers could gain dangerous authority and political influence, for the illiterate villagers in backward areas used to rely completely upon the "saint." Yet, other masters have raised their voices against social inequality and have tried, even at the cost of their lives, to change social and political conditions for the better and to spiritually revive the masses. The missionary activities of the Ṣūfīs have enlarged the fold of the faithful. The importance of Ṣūfism for spiritual education, and inculcation in the faithful of the virtues of trust in God, piety, faith in God's love, and veneration of the Prophet, cannot be overrated. The *dhikr* formulas still preserve their consoling and quieting power even for the illiterate. Mysticism permeates Persian literature and other literatures influenced by it. Such poetry has always been a source of happiness for millions, although some modernists have disdained its "narcotic" influence on Muslim thinking.

Industrialization and modern life have led to a constant decrease in the influence of Ṣūfī orders in many countries. The spiritual heritage is preserved by individuals who sometimes try to show that mystical experience conforms to modern science. Today in the West, Ṣūfism is popularized, but the genuinely and authentically devout are aware that it requires strict discipline, and that its goal can be reached—if at all—as they say, only by throwing oneself into the consuming fire of divine love. (An.Sc.)

Islāmic philosophy

Origin and inspiration of Islāmic philosophy

The origin and inspiration of philosophy in Islām are quite different from those of Islāmic theology. Philosophy developed out of and around the nonreligious practical and theoretical sciences; it recognized no theoretical limits other than those of human reason itself; and it assumed that the truth found by unaided reason does not disagree with the truth of Islām when both are properly understood. Islāmic philosophy was not a handmaid of theology. The two disciplines were related, because both followed the path of rational inquiry and distinguished themselves from traditional religious disciplines and from mysticism, which sought knowledge through practical, spiritual purification. Islāmic theology was Islāmic in the strict sense: it confined itself within the Islāmic religious community, and it remained separate from the Christian and Jewish theologies that developed in the same cultural context and used Arabic as a linguistic medium. No such separation is observable in the philosophy developed in the Islāmic cultural context and written in Arabic: Muslims, Christians, and Jews participated in it and separated themselves according to the philosophic rather than the religious doctrines they held.

THE EASTERN PHILOSOPHERS

Background and scope of philosophical interest in Islām. The background of philosophic interest in Islām is found in the earlier phases of theology. But its origin is found in the translation of Greek philosophic works. By the middle of the 9th century, there were enough translations of scientific and philosophic works from Greek, Pahlavi, and Sanskrit to show those who read them with care that scientific and philosophic inquiry was something more than a series of disputations based on what the theologians had called sound reason. Moreover, it became evident that there existed a tradition of observation, calculation, and theoretical reflection that had been pursued systematically, refined, and modified for over a millennium.

The basis of Islāmic philosophy in Greek philosophical and scientific works

The scope of this tradition was broad: it included the study of logic, the sciences of nature (including psychology and biology), the mathematical sciences (including music and astronomy), metaphysics, ethics, and politics. Each of these disciplines had a body of literature in which its principles and problems had been investigated by classical authors, whose positions had been, in turn, stated, discussed, criticized, or developed by various commentators. Islāmic philosophy emerged from its theological background when Muslim thinkers began to study this foreign tradition, became competent students of the ancient philosophers and scientists, criticized and developed their doctrines, clarified their relevance for the questions raised by the theologians, and showed what light they threw on the fundamental issues of revelation, prophecy, and the divine law.

Relation to the Muʿtazilah and interpretation of theological issues. *The teachings of al-Kindī.* Although the first Muslim philosopher, al-Kindī, who flourished in the first half of the 9th century, lived during the triumph of the Muʿtazilah of Baghdad and was connected with the ʿAbbāsid caliphs who championed the Muʿtazilah and patronized the Hellenistic sciences, there is no clear evidence that he belonged to a theological school. His writings show him to have been a diligent student of Greek and Hellenistic authors in philosophy and point to his familiarity with Indian arithmetic. His conscious, open, and unashamed acknowledgment of earlier contributions to scientific inquiry was foreign to the spirit, method, and purpose of the theologians of the time. His acquaintance with the writings of Plato and Aristotle was still incomplete and technically inadequate. He improved the Arabic translation of the "Theology of Aristotle" but made only a selective and circumspect use of it.

Al-Kindī's interest in scientific inquiry

Devoting most of his writings to questions of natural philosophy and mathematics, al-Kindī was particularly concerned with the relation between corporeal things, which are changeable, in constant flux, infinite, and as such unknowable, on the one hand, and the permanent world of forms (spiritual or secondary substances), which are not subject to flux yet to which man has no access except through things of the senses. He insisted that a purely human knowledge of all things is possible, through the use of various scientific devices, learning such things as mathematics and logic, and assimilating the contributions of earlier thinkers. The existence of a "supernatural" way to this knowledge in which all these requirements can be dispensed with was acknowledged by al-Kindī: God may choose to impart it to his prophets by cleansing and illuminating their souls and by giving them his aid, right guidance, and inspiration; and they, in turn, communicate it to ordinary men in an admirably clear, concise, and comprehensible style. This is the prophets' "divine" knowledge, characterized by a special mode of access and style of exposition. In principle, however, this very same knowledge is accessible to man without divine aid, even though "human" knowledge may lack the completeness and consummate logic of the prophets' divine message.

Reflection on the two kinds of knowledge—the human knowledge bequeathed by the ancients and the revealed knowledge expressed in the Qurʾān—led al-Kindī to pose a number of themes that became central in Islāmic philosophy: the rational–metaphorical exegesis of the Qurʾān and the Ḥadīth; the identification of God with the first being and the first cause; creation as the giving of being and as a kind of causation distinct from natural causation and Neoplatonic emanation; and the immortality of the individual soul.

The teachings of Abū Bakr ar-Rāzī. The philosopher whose principal concerns, method, and opposition to authority were inspired by the extreme Muʿtazilah was the physician Abū Bakr ar-Rāzī (flourished 9th–10th centuries). He adopted the Muʿtazilah's atomism and was intent on developing a rationally defensible theory of creation that would not require any change in God or attribute to him responsibility for the imperfection and evil prevalent in the created world. To this end, he expounded the view that there are five eternal principles—God, Soul, prime matter, infinite, or absolute, space, and unlimited, or absolute, time—and explained creation as the result of the unexpected and sudden turn of events (*faltah*). *Faltah* occurred when Soul, in her ignorance, desired matter and the good God eased her misery by allowing her to satisfy her desire and to experience the suffering of the material world, and then gave her reason to make her realize her mistake and deliver her from her union with matter, the cause of her suffering and of all evil. Ar-Rāzī claimed that he was a Platonist, that he disagreed with Aristotle, and that his views were those of the Ṣābians of Harran and the Brahmins (Hindu teachers).

The concept of the five eternal principles

Ismāʿīlī theologians became aware of the kinship between certain elements of his cosmology and their own. They disputed with him during his lifetime and continued afterward to refute his doctrines in their writings. According to their account of his doctrines, he was totally opposed to authority in matters of knowledge, believed in the progress of the arts and sciences, and held that all reasonable men are equally able to look after their own affairs, equally inspired and able to know the truth of what earlier men had taught, and equally able to improve upon it. Ismāʿīlī theologians were incensed, in particular, by his wholesale rejection of prophecy, particular revelation, and divine laws. They were likewise opposed to his criticisms of religion in general as a device employed by evil men and a kind of tyranny over men that exploits their innocence and credulity, perpetuates ignorance, and leads to conflicts and wars.

Although the fragmentary character of al-Kindī's and ar-Rāzī's surviving philosophic writings does not permit passing firm and independent judgment on their accomplishments, they tend to bear out the view of later Muslim students of philosophy that both lacked competence in the logical foundation of philosophy, were knowledgeable in some of the natural sciences but not in metaphysics, and were unable to narrow the gap that separated philosophy from the new religion, Islām.

The teachings of al-Fārābī. *Political philosophy and the study of religion.* The first philosopher to meet this challenge was al-Fārābī (flourished 9th–10th centuries). He saw that theology and the juridical study of the law were derivative phenomena that function within a framework set by the prophet as lawgiver and founder of a human community. In this community, revelation defines the opinions the members of the community must hold and the actions they must perform if they are to attain the earthly happiness of this world and the supreme happiness of the other world. Philosophy could not understand this framework of religion as long as it concerned itself almost exclusively with its truth content and confined the study of practical science to individualistic ethics and personal salvation.

The relationship of law and theology to the community

In contrast to al-Kindī and ar-Rāzī, al-Fārābī recast philosophy in a new framework analogous to that of the Islāmic religion. The sciences were organized within this philosophic framework so that logic, physics, mathematics, and metaphysics culminated in a political science whose subject matter is the investigation of happiness and how it can be realized in cities and nations. The central theme of this political science is the founder of a virtuous or excellent community. Included in this theme are views concerning the supreme rulers who follow the founder, their qualifications, and how the community must be ordered so that its members attain happiness as citizens rather than isolated human beings. Once this new philosophical framework was established, it became possible to conduct a philosophical investigation of all the elements that constituted the Islāmic community: the prophet-lawgiver, the aims of the divine laws, the legislation of beliefs as well as actions, the role of the successors to the founding legislator, the grounds of the interpretation or reform of the law, the classification of human communities according to their doctrines in addition to their size, and the critique of "ignorant" (pagan), "transgressing," "falsifying," and "erring" communities. Philosophical cosmology, psychology, and politics were blended by al-Fārābī into a political theology whose aim was to clarify the foundations of the Islāmic community and defend its reform in a direction that would promote scientific inquiry and encourage philosophers to play an active role in practical affairs.

Interpretation of Plato and Aristotle. Behind this public, or exoteric, aspect of al-Fārābī's work stood a massive body of more properly philosophic or scientific inquiries, which established his reputation among Muslims as the greatest philosophical authority after Aristotle, a great interpreter of the thought of Plato and Aristotle and their commentators, and a master to whom almost all major Muslim as well as a number of Jewish and Christian philosophers turned for a fuller understanding of the controversial, troublesome, and intricate questions of philosophy. Continuing the tradition of the Hellenistic masters of the Athenian and Alexandrian philosophical schools, al-Fārābī broadened the range of philosophical inquiry and fixed its form. He paid special attention to the study of language and its relation to logic. In his numerous commentaries on Aristotle's logical works, he expounded for the first time in Arabic the entire range of the scientific and nonscientific forms of argument and established the place of logic as an indispensable prerequisite for philosophic inquiry. His writings on natural science exposed the foundation and assumptions of Aristotle's physics and dealt with the arguments of Aristotle's opponents, both philosophers and scientists, pagan, Christian, and Muslim.

Significance of al-Fārābī in the dissemination of Greek philosophical thought

The analogy of religion and philosophy. Al-Fārābī's theological and political writings showed later Muslim philosophers the way to deal with the question of the relation between philosophy and religion and presented them with a complex set of problems that they continued to elaborate, modify, and develop in different directions. Starting with the view that religion is analogous or similar to philosophy, al-Fārābī argued that the idea of the true prophet-lawgiver ought to be the same as that of the true philosopher-king. Thus, he challenged both al-Kindī's view that prophets and philosophers have different and independent ways to the highest truth available to man and ar-Rāzī's view that philosophy is the only way to that knowledge. That a man could combine the functions of prophecy, lawgiving, philosophy, and kingship did not necessarily mean that these functions were identical; it did mean, however, that they all are legitimate subjects of philosophic inquiry. Philosophy must account for the powers, knowledge, and activities of the prophet, lawgiver, and king, which it must distinguish from and relate to those of the philosopher. The public, or political, function of philosophy was emphasized. Unlike Neoplatonism, which had for long limited itself to the Platonic teaching that the function of philosophy is to liberate the soul from the shadowy existence of the cave—in which knowledge can only be imperfectly comprehended as shadows reflecting the light of the truth beyond the cave (the world of senses)—al-Fārābī insisted with Plato that the philosopher must be forced to return to the cave, learn to talk to its inhabitants in a manner they can comprehend, and engage in actions that may improve their lot.

Impact on Ismāʿīlī theology. Although it is not always easy to know the immediate practical intentions of a philosopher, it must be remembered that in al-Fārābī's lifetime the fate of the Islāmic world was in the balance. The Sunnī caliphate's power hardly extended beyond Baghdad, and it appeared quite likely that the various Shīʿī sects, especially the Ismāʿīlīs, would finally overpower it and establish a new political order. Of all the movements in Islāmic theology, Ismāʿīlī theology was the one that was most clearly and massively penetrated by philosophy. Yet, its Neoplatonic cosmology, revolutionary background, antinomianism (antilegalism), and general expectation that divine laws were about to become superfluous with the appearance of the *qāʾim* (the imam of the "resurrection") all militated against the development of a coherent political theory to meet the practical demands of political life and present a viable practical alternative to the Sunnī caliphate. Al-Fārābī's theologico-political writ-

The use of al-Fārābī's theological–political writings to reform Ismāʿīlī thought ings helped point out this basic defect of Ismā'īlī theology. Under the Fāṭimids in Egypt (969–1171), Ismā'īlī theology modified its cosmology in the direction suggested by al-Fārābī, returned to the view that the community must continue to live under the divine law, and postponed the prospect of the abolition of divine laws and the appearance of the *qā'im* to an indefinite point in the future.

The teachings of Avicenna. *The "Oriental Philosophy."* Even more indicative of al-Fārābī's success is the fact that his writings helped produce a philosopher of the stature of Avicenna (flourished 10th–11th centuries), whose versatility, imagination, inventiveness, and prudence shaped philosophy into a powerful force that gradually penetrated Islāmic theology and mysticism and Persian poetry in eastern Islām and gave them universality and theoretical depth. His own personal philosophic views, he said, were those of the ancient sages of Greece (including the genuine views of Plato and Aristotle), which he had set forth in the "Oriental Philosophy," a book that has not survived and probably was not written or meant to be written. They were not identical with the common Peripatetic (Aristotelian) doctrines and were to be distinguished from the learning of his contemporaries, the Christian "Aristotelians" of Baghdad, which he attacked as vulgar, distorted, and falsified. His most voluminous writing, *Kitāb ash-shifā'* ("The Book of Healing"), was meant to accommodate the doctrines of other philosophers as well as hint at his own personal views, which are elaborated elsewhere in more imaginative and allegorical forms.

Distinction between essence and existence and the doctrine of creation. Avicenna had learned from certain hints in al-Fārābī that the exoteric teachings of Plato regarding "forms," "creation," and the immortality of individual souls were closer to revealed doctrines than the genuine views of Aristotle, that the doctrines of Plotinus and later Neoplatonic commentators were useful in harmonizing Aristotle's views with revealed doctrines, and that philosophy must accommodate itself to the divine law on the issue of creation and of reward and punishment in the hereafter, which presupposes some form of individual immortality. Following al-Fārābī's lead, Avicenna initiated a full-fledged inquiry into the question of being, in which he distinguished between essence and existence. He argued that the fact of existence cannot be inferred from or accounted for by the essence of existing things and that form and matter by themselves cannot interact and originate the movement of the universe or the progressive actualization of existing things. Existence must, therefore, be due to an agent-cause that necessitates, imparts, gives, or adds existence to an essence. To do so, the cause must be an existing thing and coexist with its effect. The universe consists of a chain of actual beings, each giving existence to the one below it and responsible for the existence of the rest of the chain below. Because an actual infinite is deemed impossible by Avicenna, this chain as a whole must terminate in a being that is wholly simple and one, whose essence is its very existence, and therefore is self-sufficient and not in need of something else to give it existence. Because its existence is not contingent on or necessitated by something else but is necessary and eternal in itself, it satisfies the condition of being the necessitating cause of the entire chain that constitutes the eternal world of contingent existing things.

Avicenna's investigation of being

All creation is necessarily and eternally dependent upon God. It consists of the intelligences, souls, and bodies of the heavenly spheres, each of which is eternal, and the sublunary sphere, which is also eternal, undergoing a perpetual process of generation and corruption, of the succession of form over matter, very much in the manner described by Aristotle.

The immortality of individual souls. There is, however, a significant exception to this general rule: the human rational soul. Man can affirm the existence of his soul from direct consciousness of his self (what he means when he says "I"); and he can imagine this happening even in the absence of external objects and bodily organs. This proves, according to Avicenna, that the soul is indivisible, immaterial, and incorruptible substance, not imprinted in matter, but created with the body, which it uses as an instrument. Unlike other immaterial substances (the intelligences and souls of the spheres), it is not pre-eternal but is generated, or made to exist, at the same time as the individual body, which can receive it, is formed. The composition, shape, and disposition of its body and the soul's success or failure in managing and controlling it, the formation of moral habits, and the acquisition of knowledge all contribute to its individuality and difference from other souls. Though the body is not resurrected after its corruption, the soul survives and retains all the individual characteristics, perfections or imperfections, that it achieved in its earthly existence and in this sense is rewarded or punished for its past deeds. Avicenna's claim that he has presented a philosophic proof for the immortality of generated ("created") individual souls no doubt constitutes the high point of his effort to harmonize philosophy and religious beliefs.

The doctrine of individual souls

Philosophy, religion, and mysticism. Having accounted for the more difficult issues of creation and the immortality of individual souls, Avicenna proceeded to explain the faculty of prophetic knowledge (the "sacred" intellect), revelation (imaginative representation meant to convince the multitude and improve their earthly life), miracles, and the legal and institutional arrangements (acts of worship and the regulation of personal and public life) through which the divine law achieves its end. Avicenna's explanation of almost every aspect of Islām is pursued on the basis of extensive exegesis of the Qur'ān and the Ḥadīth. The primary function of religion is to assure the happiness of the many. This practical aim of religion (which Avicenna saw in the perspective of Aristotle's practical science) enabled him to appreciate the political and moral functions of divine revelation and account for its form and content. Revealed religion, however, has a subsidiary function also—that of indicating to the few the need to pursue the kind of life and knowledge appropriate to rare individuals endowed with special gifts. These men must be dominated by the love of God to facilitate the achievement of the highest knowledge. In many places Avicenna appears to identify these men with the mystics. The identification of the philopher as a kind of mystic conveyed a new image of the philosopher as a member of the religious community who is distinguished from his coreligionists by his otherworldliness, dedicated to the inner truth of religion, and consumed by the love of God.

Prophetic, revelatory, and social knowledge

Avicenna's allegorical and mystical writings are usually called "esoteric" in the sense that they contain his personal views cast in an imaginative, symbolic form. The esoteric works must, then, be interpreted. Their interpretation must move away from the explicit doctrines contained in "exoteric" works such as the *Shifā'* and recover "the unmixed and uncorrupted truth" set forth in the "Oriental Philosophy." The "Oriental Philosophy," however, has never been available to anyone, and it is doubtful that it was written at all. This dilemma has made interpretation both difficult and rewarding for Muslim philosophers and modern scholars alike.

THE WESTERN PHILOSOPHERS

Background and characteristics of the western Muslim philosophical tradition. Andalusia (in Spain) and western North Africa contributed little of substance to Islāmic theology and philosophy until the 12th century. Legal strictures against the study of philosophy were more effective than in the east. Scientific interest was channelled into medicine, pharmacology, mathematics, astronomy, and logic. More general questions of physics and metaphysics were treated sparingly and in symbols, hints, and allegories. By the 12th century, however, the writings of al-Fārābī, Avicenna, and al-Ghazālī had found their way to the west. A philosophical tradition emerged, based primarily on the study of al-Fārābī. It was critical of Avicenna's philosophic innovations and not convinced that al-Ghazālī's critique of Avicenna touched philosophy as such, and it refused to acknowledge the position assigned by both to mysticism. The survival of philosophy in the west required extreme prudence, emphasis on its scientific character, abstention from meddling in political or religious matters, and abandonment of the hope of effecting extensive doctrinal or institutional reform.

Significance of Islāmic philosophers in the West

The teachings of Ibn Bājjah. *Theoretical science and intuitive knowledge.* Ibn Bājjah (died 1138) initiated this tradition with a radical interpretation of al-Fārābī's political philosophy that emphasized the virtues of the perfect but nonexistent city and the vices prevalent in all existing cities. He concluded that the philosopher must order his own life as a solitary individual, shun the company of nonphilosophers, reject their opinions and ways of life, and concentrate on reaching his own final goal by pursuing the theoretical sciences and achieving intuitive knowledge through contact with the Active Intelligence. The multitude live in a dark cave and see only dim shadows. Their ways of life and their imaginings and beliefs consist of layers of darkness that cannot be known through reason alone. Therefore, the divine law has been revealed to enable man to know this dark region. The philosopher's duty is to seek the light of the sun (the intellect). To do so, he must leave the cave, see all colours as they truly are and see light itself, and finally become transformed into that light. The end, then, is contact with Intelligence, not with something that transcends Intelligence (as in Plotinus, Ismāʿīlism, and mysticism), a doctrine criticized by Ibn Bājjah as the way of imagination, motivated by desire, and aiming at pleasure. Philosophy, he claimed, is the only way to the truly blessed state, which can be achieved only by going through theoretical science, even though it is higher than theoretical science.

The duty of the philosopher

Unconcern of philosophy with reform. Ibn Bājjah's cryptic style and the unfinished form in which he left most of his writings tend to highlight his departures from al-Fārābī and Avicenna. Unlike al-Fārābī, he is silent about the philosopher's duty to return to the cave and partake of the life of the city. He appears to argue that the aim of philosophy is attainable independently from the philosopher's concern with the best city and is to be achieved in solitude or, at most, in comradeship with philosophic souls. Unlike Avicenna, who prepared the way for him by clearly distinguishing between theoretical and practical science, Ibn Bājjah is concerned with practical science only insofar as it is relevant to the life of the philosopher. He is contemptuous of allegories and imaginative representations of philosophic knowledge, silent about theology, and shows no concern with improving the multitude's opinions and way of life.

The teachings of Ibn Ṭufayl. *The philosopher as a solitary individual.* In his philosophic story *Ḥayy ibn Yaqẓān* ("*Alive Son of Awake*"), the philosopher Ibn Ṭufayl (died 1185) fills gaps in the work of his predecessor Ibn Bājjah. The story communicates the secrets of Avicenna's "Oriental Philosophy" as experienced by a solitary hero, who grows up on a deserted island, learns about the things around him, acquires knowledge of the natural universe (including the heavenly bodies), and achieves the state of "annihilation" (*fanāʾ*) of the self in the divine reality. This is the apparent and traditional secret of the "Oriental Philosophy." But the hero's wisdom is still incomplete, for he knows nothing about other human beings, their way of life, or their laws. When he chances to meet one of them—a member of a religious community inhabiting a neighbouring island, who is inclined to reflect on the divine law and seek its inner, spiritual meanings and who has abandoned the society of his fellow men to devote himself to solitary meditation and worship—he does not at first recognize that he is a human being like himself, cannot communicate with him, and frightens him by his wild aspect. After learning about the doctrines and acts of worship of the religious community, he understands them as alluding to and agreeing with the truth that he had learned by his own unaided effort, and he goes as far as admitting the validity of the religion and the truthfulness of the prophet who gave it. He cannot understand, however, why the prophet communicated the truth by way of allusions, examples, and corporeal representations or why religion permits men to devote much time and effort to practical, worldly things.

Explanation of Avicenna's "Oriental Philosophy"

Concern for reform. His ignorance of the nature of most men and his compassion for them make the solitary hero insist on becoming their saviour. He persuades his companion to take him to his coreligionists and help him convert them to the naked truth by propagating among them "the secrets of wisdom." His education is completed when he fails in his endeavour. He learns the limits beyond which the multitude cannot ascend without becoming confused and unhappy. He also learns the wisdom of the divine lawgiver in addressing them in the way they can understand, enabling them to achieve limited ends through doctrines and actions suited to their abilities. The story ends with the hero taking leave of these people after apologizing to them for what he did and confessing that he is now fully convinced that they should not change their ways but remain attached to the literal sense of the divine law and obey its demands. He returns to his own island to continue his former solitary existence.

The hidden secret of Avicenna's "Oriental Philosophy." The hidden secret of Avicenna's "Oriental Philosophy" appears, then, to be that the philosopher must return to the cave, educate himself in the ways of nonphilosophers, and understand the incompatibility between philosophical life and the life of the multitude, which must be governed by religion and divine laws. Otherwise, his ignorance will lead him to actions dangerous to the well-being of both the community and philosophy. Because Ibn Ṭufayl's hero had grown up as a solitary human being, he lacks the kind of wisdom that could have enabled him to pursue philosophy in a religious community and be useful to such a community. Neither the conversion of the community to philosophy nor the philosopher's solitary life is a viable alternative.

The teachings of Averroës. *Philosophy.* To Ibn Ṭufayl's younger friend Averroës (Ibn Rushd, flourished 12th century) belongs the distinction of presenting a solution to the problem of the relation between philosophy and the Islāmic community in the west, a solution meant to be legally valid, theologically sound, and philosophically satisfactory. Here was a philosopher fully at home in what Ibn Bājjah had called the many layers of darkness. His legal training (he was a judge by profession) and his extensive knowledge of the history of the religious sciences (including theology) enabled him to speak with authority about the principles of Islāmic law and their application to theological and philosophic issues and to question the authority of al-Ghazālī and the Ashʿarīs to determine correct beliefs and right practices. He was able to examine in detail from the point of view of the divine law the respective claims of theology and philosophy to possess the best and surest way to human knowledge, to be competent to interpret the ambiguous expressions of the divine law, and to have presented convincing arguments that are theoretically tenable and practically salutary.

Solution to the problem of the relation of philosophy to the community

The divine law. The intention of the divine law, he argued, is to assure the happiness of all members of the community. This requires everyone to profess belief in the basic principles of religion as enunciated in the Qurʾān, the Ḥādith, and the *ijmāʿ* (consensus) of the learned and to perform all obligatory acts of worship. Beyond this, the only just requirement is to demand that each pursue knowledge as far as his natural capacity and makeup permit. The few who are endowed with the capacity for the highest, demonstrative knowledge are under a divine legal obligation to pursue the highest wisdom, which is philosophy, and they need not constantly adjust its certain conclusions to what theologians claim to be the correct interpretation of the divine law. Being dialecticians and rhetoricians, theologians are not in a position to determine what is and is not correct interpretation of the divine law so far as philosophers are concerned. The divine law directly authorizes philosophers to pursue its interpretation according to the best—*i.e.,* demonstrative or scientific—method, and theologians have no authority to interfere with the conduct of this activity or judge its conclusions.

The intention of the divine law

Theology. On the basis of this legal doctrine, Averroës judged the theologian al-Ghazālī's refutation of the philosophers ineffective and inappropriate because al-Ghazālī did not understand and even misrepresented the philosophers' positions and used arguments that only demonstrate his incompetence in the art of demonstration. He criticized al-Fārābī and Avicenna also for accommodating the theologians of their time and for departing from the path of the ancient philosophers merely to please the theologians. At the other extreme are the multitude for whom there are no more convincing arguments than those found in the divine law itself. Neither philosophers nor theologians are permitted to disclose to the multitude interpretations of the ambiguous verses of the Qur'ān or to confuse them with their own doubts or arguments. Finally, there are those who belong to neither the philosophers nor the multitude, either because they are naturally superior to the multitude but not endowed with the gift for philosophy or else are students in initial stages of philosophic training. For this intermediate group theology is necessary. It is an intermediate discipline that is neither strictly legal nor philosophic. It lacks their certain principles and sure methods. Therefore, theology must remain under the constant control of philosophy and the supervision of the divine law so as not to drift into taking positions that cannot be demonstrated philosophically or that are contrary to the intention of the divine law. Averroës himself composed a work on theology to show how these requirements can be met: *Kitāb al-kashf 'an manāhij al-adillah* ("Exposition of the Methods of Proofs"). In the Latin West he was best known for his philosophical answer to al-Ghazālī, *Tahāfut at-tahāfut* ("Incoherence of the Incoherence"), and for his extensive commentaries on Aristotle, works that left their impact on medieval and renaissance European thought.

The new wisdom: synthesis of philosophy and mysticism

PHILOSOPHY, TRADITIONALISM, AND THE NEW WISDOM

The role of Arabic philosophical literature in the West

Philosophy. The western tradition in Islāmic philosophy formed part of the Arabic philosophic literature that was translated into Hebrew and Latin and that played a significant role in the development of medieval philosophy in the Latin West and the emergence of modern European philosophy. Its impact on the development of philosophy in eastern Islām was not as dramatic, but was important nevertheless. Students of this tradition—*e.g.*, the prominent Jewish philosopher Maimonides (flourished 12th century) and the historian Ibn Khaldūn (flourished 14th century)—moved to Egypt, where they taught and had numerous disciples. Most of the writings of Ibn Bājjah, Ibn Ṭufayl, and Averroës found their way to the east also, where they were studied alongside the writings of their eastern predecessors. In both regions thinkers who held to the idea of philosophy as formulated by the eastern and western philosophers thus far discussed continued to teach. They became isolated and overwhelmed, however, by the resurgence of traditionalism and the emergence of a new kind of philosophy whose champions looked on the earlier masters as men who had made significant contributions to the progress of knowledge but whose overall view was defective and had now become outdated.

Traditionalism and the new wisdom. Resurgent traditionalism found effective defenders in men such as Ibn Taymīyah (13th–14th centuries) who employed a massive battery of philosophic, theological, and legal arguments against every shade of innovation and called for a return to the beliefs and practices of the pious ancestors. These attacks, however, did not deal a decisive blow to philosophy as such. It rather drove philosophy underground for a period, only to re-emerge in a new garb. A more important reason for the decline of the earlier philosophic tradition, however, was the renewed vitality and success of the program formulated by al-Ghazālī for the integration of theology, philosophy, and mysticism into a new kind of philosophy called wisdom (*ḥikmah*). It consisted of a critical review of the philosophy of Avicenna, preserving its main external features (its logical, physical, and, in part, metaphysical structure, and its terminology) and introducing principles of explanation for the universe and its relation to God based on personal experience and direct vision.

The emergence of philosophy under a new garb

Characteristic features of the new wisdom. If the popular theology preached by the philosophers from al-Fārābī to Averroës is disregarded, it is evident that philosophy proper meant to them what al-Fārābī called a state of mind dedicated to the quest and the love for the highest wisdom. None of them claimed, however, that he had achieved this highest wisdom. In contrast, every leading exponent of the new wisdom stated that he had achieved or received it through a private illumination, dream (at times inspired by the Prophet), or vision and on this basis proceeded to give an explanation of the inner structure of natural and divine things. In every case, this explanation incorporated Platonic or Aristotelian elements but was more akin to some version of a later Hellenistic philosophy, which had found its way earlier into one or another of the schools of Islāmic theology, though, because of the absence of an adequate philosophic education on the part of earlier theologians, it had not been either elaborated or integrated into a comprehensive view. Like their late-Hellenistic counterparts, exponents of the new wisdom proceeded through an examination of the positions of Plato, Aristotle, and Plotinus. They also gave special attention to the insights of the pre-Socratic philosophers of ancient Greece and the myths and revelations of the ancient Near East, and they offered to resolve the fundamental questions that had puzzled earlier philosophers. In its basic movement and general direction, therefore, Islāmic philosophy between the 9th and the 19th centuries followed a course parallel to that of Greek philosophy from the 5th century BC to the 6th century AD.

Use of Greek philosophers by exponents of the new wisdom

Critiques of Aristotle in Islāmic theology. The critique of Aristotle that had begun in Mu'tazilī circles and had found a prominent champion in Abū Bakr ar-Rāzī was provided with a more solid foundation in the 10th and 11th centuries by the Christian theologians and philosophers of Baghdad, who translated the writings of the Hellenistic critics of Aristotle (*e.g.*, John Philoponus) and made use of their arguments in commenting on Aristotle and in independent theological and philosophic works. Avicenna's attack on these so-called Aristotelians and their Hellenistic predecessors (an attack that had been initiated by al-Fārābī and was to be continued by Averroës) did not prevent the spread of their theologically based anti-Aristotelianism among Jewish and Muslim students of philosophy in the 12th century, such as Abū al-Barakāt al-Baghdādī (died *c.* 1175) and Fakhr ad-Dīn ar-Rāzī. These theologians continued and intensified al-Ghazālī's attacks on Avicenna and Aristotle (especially their views on time, movement, matter, and form, the nature of the heavenly bodies, and the relation between the intelligible and sensible worlds). They suggested that a thorough examination of Aristotle had revealed to them, on philosophic grounds, that the fundamental disagreements between him and the theologies based on the revealed religions represented open options and that Aristotle's view of the universe was in need of explanatory principles that could very well be supplied by theology. This critique provided the framework for the integration of philosophy into theology from the 13th century onward.

Synthesis of philosophy and mysticism. Although it made use of such theological criticisms of philosophy, the new wisdom took the position that theology did not offer a positive substitute for and was incapable of solving the difficulties of "Aristotelian" philosophy. It did not question the need to have recourse to the Qur'ān and the Ḥadīth to find the right answers. It insisted (on the authority of a long-standing mystical tradition), however, that theology concerns itself only with the external expressions of this divine source of knowledge. The inner core was reserved for the adepts of the mystic path whose journey leads to the experience of the highest reality in dreams and visions. Only the mystical adepts are in possession of the one true wisdom, the ground of both the external expressions of the divine law and the phenomenal world of human experience and thought.

The inner core of divine knowledge as reserved for the adepts

PRIMARY TEACHERS OF THE NEW WISDOM

The teachings of as-Suhrawardī. The first master of the new wisdom, as-Suhrawardī (12th century), called it the "Wisdom of Illumination." He rejected Avicenna's distinction between essence and existence and Aristotle's distinction between substance and accidents, possibility and actuality, and matter and form, on the ground that they are mere distinctions of reason. Instead, he concentrated on the notion of being and its negation, which he called "light" and "darkness," and explained the gradation of beings as gradation of their mixture according to the degree of "strength," or "perfection," of their light. This gradation forms a single continuum that culminates in pure light, self-luminosity, self-awareness, self-manifestation, or self-knowledge, which is God, the light of lights, the true One. The stability and eternity of this single continuum result from every higher light overpowering and subjugating the lower, and movement and change in it result from each of the lower lights desiring and loving the higher.

As-Suhrawardī's "pan-lightism" is not particularly close to traditional Islāmic views concerning the creation of the world and God's knowledge of particulars. The structure of his universe remains largely that of the Platonists and the Aristotelians. And his account of the emanation process avoids the many difficulties that had puzzled Neoplatonists as they tried to understand how the second hypostasis (reality) proceeds from the One. He asserted that it proceeds without in any way affecting the One and that the One's self-sufficiency is enough to explain the giving out that seems to be both spontaneous and necessary. His doctrine is presented in a way that suggests that it is the inner truth behind the exoteric (external) teachings of Islām as well as Zoroastrianism, indeed the wisdom of all ancient sages, especially Iranians and Greeks, and the revealed religions as well. This neutral yet positive attitude toward the diversity of religions, which was not absent among Muslim philosophers and mystics, was to become one of the hallmarks of the new wisdom. Different religions were seen as different manifestations of the same truth, their essential agreement was emphasized, and various attempts were made to combine them into a single harmonious religion meant for all of mankind.

As-Suhrawardī takes an important step in this direction through his doctrine of imaginative-bodily "resurrection." After their departure from the prison of the body, souls that are fully purified ascend directly to the world of separate lights. The ones that are only partially purified or are evil souls escape to a "world of images" suspended below the higher lights and above the corporeal world. In this world of images, or forms (not to be confused with the Platonic forms, which as-Suhrawardī identifies with higher and permanent intelligible lights), partially purified souls remain suspended and are able to create for themselves and by their own power of imagination pleasing figures and desirable objects in forms more excellent than their earthly counterparts and are able to enjoy them forever. Evil souls become dark shadows, suffer (presumably because their corrupt and inefficient power of imagination can create only ugly and frightening forms), and wander about as ghosts, demons, and devils. The creative power of the imagination, which as a human psychological phenomenon was already used by the philosophers to explain prophetic powers, was seized upon by the new wisdom as "divine magic." It was used to construct an eschatology, to explain miracles, dreams, and other saintly theurgic (healing) practices, to facilitate the movement between various orders of being, and for literary purposes.

The teachings of Ibn al-ʿArabī. The account of the doctrines of Ibn al-ʿArabī (12th–13th centuries) belongs properly to the history of Islāmic mysticism. Yet his impact on the subsequent development of the new wisdom was in many ways far greater than was that of as-Suhrawardī. This is true especially of his central doctrine of the "unity of being" and his sharp distinction between the absolute One, which is undefinable Truth (*ḥaqq*), and his self-manifestation (*ẓuhūr*), or creation (*khalq*), which is ever new (*jadīd*) and in perpetual movement, a movement that unites the whole of creation in a process of constant renewal. At the very core of this dynamic edifice stands nature, the "dark cloud" (*ʿamā'*) or "mist" (*bukhār*), as the ultimate principle of things and forms: intelligence, heavenly bodies, and elements and their mixtures that culminate in the "perfect man." This primordial nature is the "breath" of the Merciful God in his aspect as Lord. It "flows" throughout the universe and manifests Truth in all its parts. It is the first mother through which Truth manifests itself to itself and generates the universe. And it is the universal natural body that gives birth to the translucent bodies of the spheres, to the elements, and to their mixtures, all of which are related to that primary source as daughters to their mother.

Doctrine of the "unity of being"

Ibn al-ʿArabī attempted to explain how Intelligence proceeds from the absolute One by inserting between them a primordial feminine principle, which is all things in potentiality but which also possesses the capacity, readiness, and desire to manifest or generate them first as archetypes in Intelligence and then as actually existing things in the universe below. Ibn al-ʿArabī gave this principle numerous names, including prime "matter" (*ʿunṣur*), and characterized it as the principle "whose existence makes manifest the essences of the potential worlds." The doctrine that the first simple originated thing is not Intelligence but "indefinite matter" and that Intelligence was originated through the mediation of this matter was attributed to Empedocles, a 5th-century-BC Greek philosopher, in doxographies (compilations of extracts from the Greek philosophers) translated into Arabic. It represented an attempt to bridge the gulf between the absolute One and the multiplicity of forms in Intelligence. The Andalusian mystic Ibn Masarrah (9th–10th centuries) is reported to have championed pseudo-Empedoclean doctrines, and Ibn al-ʿArabī (who studied under some of his followers) quotes Ibn Masarrah on a number of occasions. This philosophic tradition is distinct from the one followed by the Ismāʿīlī theologians, who explained the origination of Intelligence by the mediation of God's will.

The teachings of Twelver Shīʿism and the school of Eṣfahān. After Ibn al-ʿArabī, the new wisdom developed rapidly in intellectual circles in eastern Islām. Commentators on the works of Avicenna, as-Suhrawardī, and Ibn al-ʿArabī began the process of harmonizing and integrating the views of the masters. Great poets made them part of every educated man's literary culture. Mystical fraternities became the custodians of such works, spreading them into Central Asia and the Indian subcontinent and transmitting them from one generation to another. Following the Mongol khan Hülagü's entry into Baghdad (1258), the Twelver Shīʿah were encouraged by the Il Khanid Tatars and Naṣīr ad-Dīn aṭ-Ṭūsī (the philosopher and theologian who accompanied Hülagü as his vizier) to abandon their hostility to mysticism. Muʿtazilī doctrines were retained in their theology. Theology, however, was downgraded to "formal" learning that must be supplemented by higher things, the latter including philosophy and mysticism, both of earlier Shīʿī (including Ismāʿīlī) origin and of later Sunnī provenance. Al-Ghazālī, as-Suhrawardī, Ibn al-ʿArabī, and Avicenna were then eagerly studied and (except for their doctrine of the imamate) embraced with little or no reservation. This movement in Shīʿī thought gathered momentum when the leaders of a mystical fraternity established themselves as the Ṣafavid dynasty (1501–1732) in Iran, where they championed Twelver Shīʿism as the official doctrine of the new monarchy. During the 17th century, Iran experienced a cultural and scientific renaissance that included a revival of philosophic studies. There, Islāmic philosophy found its last creative exponents. The new wisdom as expounded by the masters of the school of Eṣfahān radiated throughout eastern Islām and continued as a vital tradition until modern times.

The role of mystical fraternities in the east

The major figures of the school of Eṣfahān were Mīr Dāmād (Muḥammad Bāqir ibn ad-Dāmād, died 1631/32) and his great disciple Mullā Ṣadrā (Ṣadr ad-Dīn ash-Shīrāzī, *c.* 1571–1640). Both were men of wide culture and prolific writers with a sharp sense for the history and development of philosophic ideas.

The teachings of Mīr Dāmād. Mīr Dāmād was the first to expound the notion of "eternal origination" (*ḥudūth dahrī*) as an explanation for the creation of the world. Muslim philosophers and their critics had recognized the crucial role played by the question of time in the discussion of the eternity of the world. The proposition that time is the measure of movement was criticized by Abū al-Barakāt al-Baghdādī, who argued that time is prior to movement and rest, indeed to everything except being. Time is the measure or concomitant of being, lasting and transient, enduring and in movement or rest. It character-

The concept of "eternal origination"

izes or qualifies all being, including God. God works in time, incessantly willing and directly creating everything in the world: his persistent will creates the eternal beings of the world, and his ever-renewed will creates the transient beings. The notion of a God who works in time was of course objectionable to theology, and Fakhr ad-Dīn ar-Rāzī refused to accept this solution despite its attractions. Ar-Rāzī also saw that it leads to the notion (attributed to Plato) that time is a self-subsistent substance, whose relation to God would further compromise his unity. Finally, ar-Rāzī explained that this self-subsistent substance will have to be related to different beings in different ways. It is called "everlastingness" (*sarmad*) when related to God and the Intelligences (angels) that are permanent and do not move or change in any way, "eternity" (*dahr*) when related to the totality of the world of movement and change, and "time" (*zamān*) when related to corporeal beings that make up the world of movement and change.

The problems of essence and existence

Mīr Dāmād returned to Avicenna and sought to harmonize his views with those of as-Suhrawardī on the assumption that what Avicenna meant by his "Oriental" (*mashriqīyah*) philosophy was identical with as-Suhrawardī's wisdom of "illumination" (*ishrāq*), which he interpreted as a Platonic doctrine that asserted the priority of essence (form) over being (existence). Time, for Mīr Dāmād, was neither a mere being of reason nor an accident of existing things. It belongs to the essence of things and describes their mode and rank of being. It is a "relation" that beings have to each other because of their essential nature. There must, therefore, be three ranks of order of time corresponding to the three ranks of order of being. Considered as the relation of God to the divine names and attributes (Intelligences or archetypes), the relation is "everlastingness." Considered as the relation between the Intelligences, or archetypes, and their reflections in the mutable things of the world below, the relation is "eternity." And considered as the relation between these mutable things, the relation is "time." Creation, or origination, is this very relation. Thus, the origination of the immutable Intelligences, or archetypes, is called "everlasting creation," the origination of the world of mutable beings as a whole is called "eternal creation," and the generation of mutable things within the world is called "temporal creation."

The teachings of Mullā Ṣadrā. Mullā Ṣadrā superimposed Ibn al-ʿArabī's mystical thought (whose philosophic implications had already been exposed by a number of commentators) on the "Aristotelian"–Illuminationist synthesis developed by Mīr Dāmād. Against his master, he argued with the Aristotelians for the priority of being (existence) over essence (form), which he called an abstraction; and, with Ibn al-ʿArabī, he argued for the "unity of being" within which beings differ only according to "priority and posteriority," "perfection and imperfection," and "strength and weakness." All being is thus viewed as a graded manifestation, or determination, of absolute, or pure, Being, and every level of being possesses all the attributes of pure Being, but with varying degrees of intensity or perfection.

Mullā Ṣadrā's doctrine of nature

Mullā Ṣadrā considered his unique contribution to Islāmic philosophy to be his doctrine of nature, which enabled him to assert that everything other than God and his knowledge—*i.e.*, the entire corporeal world, including the heavenly bodies—is originated "eternally" as well as "temporally." This doctrine of nature is an elaboration of the last manifestation of Ibn al-ʿArabī's "nature" or prime "matter," articulated on philosophic grounds and within the general framework of Aristotelian natural science and defended against every possible philosophic and theological objection.

Nature for Mullā Ṣadrā is the "substance" and "power" of all corporeal beings and the direct cause of their movement. Movement (and time, which measures it) is therefore not an accident of substance or an accompaniment of some of its accidents. It signifies the very change, renewal, and passing of being—itself being in constant "flow," or flux. The entire corporeal world, both the celestial spheres and the world of the elements, constantly renews itself. The "matter" of corporeal things has the power to become a new form at every instant; and the resulting matter–form complex is at every instant a new matter ready for, desiring, and moving toward another form. Men fail to observe this constant flux and movement in simple bodies not because of the endurance of the same form in them but because of the close similarity between their ever-new forms. What the philosophers call "movement" and "time" are not, as they believed, anchored in anything permanent—*e.g.*, in what they call "nature," "substance," or "essence"; essence is permanent only in the mind, and nature and substance are permanent activity. Nature as permanent activity is the very being of natural things and identical with their substance. Because nature is "permanent" in this sense, it is connected to a permanent principle that manifests activity in it permanently. Because nature constantly renews itself, all renewed and emergent things are connected to it. Thus, nature is the link between what is eternal and what is originated, and the world of nature is originated both eternally and temporarily.

The problem of movement, change, or renewal

Mullā Ṣadrā distinguishes this primary "movement-in-substance" (*al-ḥarakah fī al-jawhar*) from haphazard, compulsory, and other accidental movements that lack proper direction, impede the natural movement of substance, or reverse it. Movement-in-substance is not universal change or flux without direction, the product of conflict between two equally powerful principles, or a reflection of the nonbeing of the world of nature when measured against the world of permanent forms. It is, rather, the natural beings' innate desire to become more perfect, which directs this ceaseless self-renewal, self-origination, or self-emergence into a perpetual and irreversible flow upward in the scale of being—from the simplest elements to the human body–soul complex and the heavenly body–soul complex (both of which participate in the general instability, origination, and passing of being that characterizes the entire corporeal world). This flow upward, however, is by no means the end. For the indefinite "matter" (Ibn al-ʿArabī's "cloud" and the mystics' "created Truth") is the "substratum" of everything other than its Creator, the mysterious pure Truth. It "extends" beyond the body–soul complex to the Intelligences (divine names) that are Being's first, highest, and purest actualization or activity. This "extension" unites everything other than the Creator into a single continuum. The human body–soul complex and the heavenly body–soul complex are not moved externally by the Intelligences. Their movement is an extension of the process of self-perfection. Having reached the highest rank of order of substance in the corporeal world, they are now prepared, and still moved by their innate desire, to flow upward and transform themselves into pure intelligence.

IMPACT OF MODERNISM

The new wisdom lived on during the 18th and 19th centuries, conserving much of its vitality and strength but not cultivating new ground. It attracted able thinkers such as Shāh Walī Allāh of Delhi and Hādī Sabzevārī and became a regular part of the program of higher education in the cultural centres of the Ottoman Empire, Iran, and the Indian subcontinent, a status never achieved by the earlier tradition of Islāmic philosophy. In collaboration with its close ally Persian mystical poetry, the new wisdom determined the intellectual outlook and spiritual mood of educated Muslims in the regions where Persian had become the dominant literary language.

The problem of reform in modern Islāmic thought

The wholesale rejection of the new wisdom in the name of simple, robust, and more practical piety (which had been initiated by Ibn Taymīyah and which continued to find exponents among jurists) made little impression on its devotees. To be taken seriously, reform had to come from their own ranks and be espoused by such thinkers as the eminent theologian and mystic of Muslim India Aḥmad Sirhindī (flourished 16th–17th centuries)—a reformer who spoke their language and attacked Ibn al-ʿArabī's "unity of being" only to defend an older, presumably more orthodox form of mysticism. Despite some impact, however, attempts of this kind remained isolated and were either ignored or reintegrated into the mainstream, until the coming of the modern reformers. The 19th- and 20th-century reformers Jamāl ad-Dīn al-Afghānī, Muḥammad

ʿAbduh, and Muḥammad Iqbāl were initially educated in this tradition, but they rebelled against it and advocated radical reforms.

The modernists attacked the new wisdom at its weakest point; that is, its social and political norms, its individualistic ethics, and its inability to speak intelligently about social, cultural, and political problems generated by a long period of intellectual isolation that was further complicated by the domination of the European powers. Unlike the earlier tradition of Islāmic philosophy from al-Fārābī to Averroës, which had consciously cultivated political science and investigated the political dimension of philosophy and religion and the relation between philosophy and the community at large, the new wisdom from its inception lacked genuine interest in these questions, had no appreciation for political philosophy, and had only a benign toleration for the affairs of the world.

None of the reformers was a great political philosopher. They were concerned with reviving their nations' latent energies, urging them to free themselves from foreign domination, and impressing on them the need to reform their social and educational institutions. They also saw that all this required a total reorientation, which could not take place so long as the new wisdom remained not only the highest aim of a few solitary individuals but also a social and popular ideal as well. Yet, as late as 1917, Iqbāl found that "the present-day Muslim prefers to roam about aimlessly in the valley of Hellenic-Persian mysticism, which teaches us to shut our eyes to the hard reality around, and to fix our gaze on what is described as 'illumination.' " His reaction was harsh: "To me this self-mystification, this nihilism, *i.e.*, seeking reality where it does not exist, is a physiological symptom, giving me a clue to the decadence of the Muslim world."

To arrest the decadence and infuse new vitality in a society in which they were convinced religion must remain the focal point, the modern reformers advocated a return to the movements and masters of Islāmic theology and philosophy antedating the new wisdom. They argued that these, rather than the "Persian incrustation of Islām," represented Islām's original and creative impulse. The modernists were attracted, in particular, to the views of the Muʿtazilah: affirmation of God's unity and denial of all similarity between him and created things; reliance on human reason; emphasis on man's freedom; faith in man's ability to distinguish between good and bad; and insistence on man's responsibility to do good and fight against evil in private and public places. They were also impressed by the traditionalists' devotion to the original, uncomplicated forms of Islām and by their fighting spirit, and by the Ashʿarīs' view of faith as an affair of the heart and their spirited defense of the Muslim community. In viewing the scientific and philosophic tradition of eastern and western Islām prior to the Tatar and Mongol invasions, they saw an irrefutable proof that true Islām stands for the liberation of man's spirit, promotes critical thought, and provides both the impetus to grapple with the temporal and the demonstration of how to set it in order. These ideas initiated what was to become a vast effort to recover, edit, and translate into the Muslim national languages works of earlier theologians and philosophers, which had been long neglected or known only indirectly through later accounts.

The call to return to the early masters of Islāmic theology and philosophy

The modern reformers insisted, finally, that Muslims must be taught to understand the real meaning of what has happened in Europe, which in effect means the understanding of modern science and philosophy, including modern social and political philosophies. Initially, this challenge became the task of the new universities in the Muslim world. In the latter part of the 20th century, however, the originally wide gap between the various programs of theological and philosophic studies in religious colleges and in modern universities narrowed considerably.

The role of universities in the reform of Islāmic theology and philosophy

(M.S.M./Ed.)

THE CULTURE OF ISLĀM

Islāmic Law, Sharīʿah

Total and unqualified submission to the will of Allāh (God) is the fundamental tenet of Islām: Islāmic law is therefore the expression of Allāh's command for Muslim society and, in application, constitutes a system of duties that are incumbent upon a Muslim by virtue of his religious belief. Known as the Sharīʿah (literally, "the path leading to the watering place"), the law constitutes a divinely ordained path of conduct that guides the Muslim toward a practical expression of his religious conviction in this world and the goal of divine favour in the world to come.

NATURE AND SIGNIFICANCE OF ISLĀMIC LAW

Muslim jurisprudence, the science of ascertaining the precise terms of the Sharīʿah, is known as *fiqh* (literally "understanding"). The historical process of the discovery of Allāh's law (see below) was regarded as completed by the end of the 9th century when the law had achieved a definitive formulation in a number of legal manuals written by different jurists. Throughout the medieval period this basic doctrine was elaborated and systematized in a large number of commentaries, and the voluminous literature thus produced constitutes the traditional textual authority of Sharīʿah law.

Scope of Sharīʿah

In classical form the Sharīʿah differs from Western systems of law in two principal respects. In the first place the scope of the Sharīʿah is much wider, since it regulates man's relationship not only with his neighbours and with the state, which is the limit of most other legal systems, but also with his God and his own conscience. Ritual practices, such as the daily prayers, almsgiving, fasting, and pilgrimage, are an integral part of Sharīʿah law and usually occupy the first chapters in the legal manuals. The Sharīʿah is also concerned as much with ethical standards as with legal rules, indicating not only what man is entitled or bound to do in law, but also what he ought, in conscience, to do or refrain from doing. Accordingly, certain acts are classified as praiseworthy (*mandūb*), which means that their performance brings divine favour and their omission divine disfavour, and others as blameworthy (*makrūh*), which means that omission brings divine favour and commission divine disfavour; but in neither case is there any legal sanction of punishment or reward, nullity or validity. The Sharīʿah is not merely a system of law, but a comprehensive code of behaviour that embraces both private and public activities.

The second major distinction between the Sharīʿah and Western legal systems is the result of the Islāmic concept of law as the expression of the divine will. With the death of the Prophet Muḥammad in 632, communication of the divine will to man ceased so that the terms of the divine revelation were henceforth fixed and immutable. When, therefore, the process of interpretation and expansion of this source material was held to be complete with the crystallization of the doctrine in the medieval legal manuals, Sharīʿah law became a rigid and static system. Unlike secular legal systems that grow out of society and change with the changing circumstances of society, Sharīʿah law was imposed upon society from above. In Islāmic jurisprudence it is not society that moulds and fashions the law, but the law that precedes and controls society.

Such a philosophy of law clearly poses fundamental problems of principle for social advancement in contemporary Islām. How can the traditional Sharīʿah law be adapted to meet the changing circumstances of modern Muslim society? This is now the central issue in Islāmic law. (See below *Reform of Sharīʿah law*).

HISTORICAL DEVELOPMENT OF SHARĪʿAH LAW

For the first Muslim community established under the leadership of the Prophet at Medina in 622, the Qurʾānic revelations laid down basic standards of conduct. But the Qurʾān is in no sense a comprehensive legal code. No more than 80 verses deal with strictly legal matters; while these verses cover a wide variety of topics and introduce many novel rules, their general effect is simply to modify the existing Arabian customary law in certain important particulars.

During his lifetime Muḥammad, as the supreme judge of the community, resolved legal problems as they arose by interpreting and expanding the general provisions of the Qur'ān, and the same *ad hoc* activity was carried on after his death by the caliphs (temporal and spiritual rulers) of Medina. But the foundation of the Umayyad dynasty in 661, governing from its centre of Damascus a vast military empire, produced a legal development of much broader dimensions. With the appointment of judges, or *qāḍīs*, to the various provinces and districts, an organized judiciary came into being. The *qāḍīs* were responsible for giving effect to a growing corpus of Umayyad administrative and fiscal law; and since they regarded themselves essentially as the spokesmen of the local law, elements and institutions of Roman-Byzantine and Persian-Sāsānian law were absorbed into Islāmic legal practice in the conquered territories. Depending upon the discretion of the individual *qāḍī*, decisions would be based upon the rules of the Qur'ān where these were relevant; but the sharp focus in which the Qur'ānic laws were held in the Medinian period had become lost with the expanding horizons of activity.

Development of a judiciary

Development of different schools of law. A reaction to this situation arose in the early 8th century when pious scholars, grouped together in loose, studious fraternities, began to debate whether or not Umayyad legal practice was properly implementing the religious ethic of Islām. Actively sponsored by the 'Abbāsid rulers, who came to power in the mid-8th century pledged to build a truly Islāmic state and society, the activities of the jurists (*faqīh*, plural *fuqahā'*) in these early schools of law marked the real beginning of Islāmic jurisprudence. Their aim was to Islāmize the law by reviewing the current legal practice in the light of the Qur'ānic principles and then on this basis adopting, modifying, or rejecting the practice as part of their ideal scheme of law.

Of the many early schools of law the two most important were those of the Mālikīs in Medina and the Ḥanafīs in al-Kūfah, named after two outstanding scholars in the respective localities, Mālik ibn Anas and Abū Ḥanīfah. Inevitably the Mālikī and Ḥanafī doctrines, as they were then being recorded in the first compendiums of law, differed considerably from each other, not only because free juristic speculation was bound to produce varying results but also because the thought of the scholars was conditioned by their different social environments. A deep conflict of juristic principle emerged within the schools between those who maintained that outside the terms of the Qur'ān scholars were free to use their reason (*ra'y*) to ascertain the law and those who insisted that the only valid source of law outside the Qur'ān lay in the precedents set by the Prophet himself.

The jurist ash-Shāfi'ī (died 820) aimed to eliminate these schisms and produce greater uniformity in the law by expounding a firm theory of the sources from which the law must be derived. Ash-Shāfi'ī's fundamental teaching was that knowledge of the Sharī'ah could be attained only through divine revelation found either in the Qur'ān or in the divinely inspired traditions (*sunnah*) of the Prophet as ascertained through authentic reports (Ḥadīth). Human reason in law should be strictly confined to the process of analogical deduction, or *qiyās*—problems not specifically answered by the divine revelation were to be solved by applying the principles upon which closely parallel cases had been regulated by the Qur'ān or *sunnah*.

The role of divinely inspired traditions

Shāfi'ī's insistence upon the importance of the *sunnah* as a source of law produced a great activity in the collection and classification of Ḥadīths, particularly among his own supporters, who formed the Shāfi'ī school, and the followers of Aḥmad ibn Ḥanbal (died 855) who formed the Ḥanbalī school. Muslim scholarship maintained that the classical compilations of Ḥadīths—especially those of Bukhārī (died 870) and Muslim (died 875)—constituted an authentic record of the Prophet's precedents. The general view of Western orientalists, however, is that a considerable part of the *sunnah* represents the views of later jurists fictitiously ascribed to the Prophet to give the doctrine a greater authority.

Later developments. Shāfi'ī's thesis formed the basis of the classical theory of the roots of jurisprudence (*uṣūl al-fiqh*), which crystallized in the early 10th century. Juristic "effort" to comprehend the terms of the Sharī'ah is known as *ijtihād*, and legal theory first defines the course that *ijtihād* must follow. In seeking the answer to a legal problem the jurist must first consult the Qur'ān and the *sunnah*. Failing any specific solution in this divine revelation he must employ analogy (*qiyās*) or certain subsidiary principles of reasoning—*istiḥsān* (equitable preference) and *istiṣlāḥ* (the public interest). The legal theory then evaluates the results of *ijtihād* on the basis of the criterion of *ijmā'* (consensus). As an attempt to define Allāh's law, the *ijtihād* of individual scholars could result only in a tentative conclusion termed *ẓann* ("conjecture"). But where a conclusion became the subject of unanimous agreement by the qualified scholars, it became a certain (*yaqīn*) and infallible expression of Allāh's law.

Two major effects flowed from this classical doctrine of *ijmā'*. It served first as a permissive principle to admit the validity of variant opinions as equally probable attempts to define the Sharī'ah. Second, it operated as a restrictive principle to ratify the status quo; for once the *ijmā'* had cast an umbrella authority not only over those points that were the subject of a consensus but also over existing variant opinions, to propound any further variant was to contradict the infallible *ijmā'* and therefore tantamount to heresy.

Ijmā' set the final seal of rigidity upon the doctrine, and from the 10th century onward independent juristic speculation ceased. In the Arabic expression, "the door of *ijtihād* was closed." Henceforth jurists were *muqallids*, or imitators, bound by the doctrine of *taqlīd* ("clothing with authority" *i.e.,* unquestioned acceptance) to follow the doctrine as it was recorded in the authoritative legal manuals.

Sharī'ah law is a candidly pluralistic system, the philosophy of the equal authority of the different schools being expressed in the alleged dictum of the Prophet: "Difference of opinion among my community is a sign of the bounty of Allāh." But outside the four schools of Sunnī, or orthodox, Islām stand the minority sects of the Shī'ah and the Ibāḍīs whose own versions of the Sharī'ah differ considerably from those of the Sunnīs. Shī'ī law in particular grew out of a fundamentally different politico-religious system in which the rulers, or *imāms*, were held to be divinely inspired and therefore the spokesmen of the Lawgiver himself. Geographically, the division between the various schools and sects became fairly well defined as the *qāḍīs*' courts in different areas became wedded to the doctrine of one particular school. Thus Ḥanafī law came to predominate in the Middle East and the Indian subcontinent; Mālikī law in North, West, and Central Africa; Shāfi'ī law in East Africa, the southern parts of the Arabian peninsula, Malaysia, and Indonesia; Ḥanbalī law in Saudi Arabia, Shī'ī law in Iran and the Shī'ī communities of India and East Africa; Ibāḍī law in Zanzibar, 'Umān, and parts of Algeria.

Distribution of various schools of Islāmic law

Although Sharī'ah doctrine was all-embracing, Islāmic legal practice has always recognized jurisdictions other than that of the *qāḍīs*. Because the *qāḍīs*' courts were hidebound by a cumbersome system of procedure and evidence, they did not prove a satisfactory organ for the administration of justice in all respects, particularly as regards criminal, land, and commercial law. Hence, under the broad head of the sovereign's administrative power (*siyāsah*), competence in these spheres was granted to other courts, known collectively as *maẓālim* courts, and the jurisdiction of the *qāḍīs* was generally confined to private family and civil law. As the expression of a religious ideal, Sharī'ah doctrine was always the focal point of legal activity, but it never formed a complete or exclusively authoritative expression of the laws that in practice governed the lives of Muslims.

THE SUBSTANCE OF TRADITIONAL SHARĪ'AH LAW

Sharī'ah duties are broadly divided into those that an individual owes to Allāh (the ritual practices or *'ibādāt*) and those that he owes to his fellow men (*mu'āmalāt*). It is the latter category of duties alone, constituting law in the Western sense, that is described here.

Penal law. Offenses against the person, from homicide to assault, are punishable by retaliation (*qiṣāṣ*), the offender being subject to precisely the same treatment as his victim. But this type of offense is regarded as a civil injury rather than a crime in the technical sense, since it is not the state but only the victim or his family who have the right to prosecute and to opt for compensation or blood money (*diyah*) in place of retaliation.

Fixed punishment

For six specific crimes the punishment is fixed (*ḥadd*): death for apostasy and for highway robbery; amputation of the hand for theft; death by stoning for extramarital sex relations (*zinā*) where the offender is a married person and 100 lashes for unmarried offenders; 80 lashes for an unproved accusation of unchastity (*qadhf*) and for the drinking of any intoxicant.

Outside the *ḥadd* crimes, both the determination of offenses and the punishment therefore lies with the discretion of the executive or the courts.

Law of transactions. A legal capacity to transact belongs to any person "of prudent judgment" (*rāshid*), a quality that is normally deemed to arrive with physical maturity or puberty. There is an irrebuttable presumption of law (1) that boys below the age of 12 and girls below the age of 9 have not attained puberty, and (2) that puberty has been attained by the age of 15 for both sexes. Persons who are not *rāshid*, on account of minority, mental deficiency, simplicity, or prodigality, are placed under interdiction: their affairs are managed by a guardian and they cannot transact effectively without the guardian's consent.

The basic principles of the law are laid down in the four root transactions of (1) sale (*bay'*), transfer of the ownership or corpus of property for a consideration; (2) hire (*ijārah*), transfer of the usufruct (right to use) of property for a consideration; (3) gift (*hibah*), gratuitous transfer of the corpus of property, and (4) loan (*'āriyah*), gratuitous transfer of the usufruct of property. These basic principles are then applied to the various specific transactions of, for example, pledge, deposit, guarantee, agency, assignment, land tenancy, partnership, and *waqf* foundations. *Waqf* is a peculiarly Islāmic institution whereby the founder relinquishes his ownership of real property, which belongs henceforth to Allāh, and dedicates the income or usufruct of the property in perpetuity to some pious or charitable purpose, which may include settlements in favour of the founder's own family.

The Islāmic law of transactions as a whole is dominated by the doctrine of *ribā*. Basically, this is the prohibition of usury, but the notion of *ribā* was rigorously extended to cover, and therefore preclude, any form of interest on a capital loan or investment. And since this doctrine was coupled with the general prohibition on gambling transactions, Islāmic law does not, in general, permit any kind of speculative transaction the results of which, in terms of the material benefits accruing to the parties, cannot be precisely forecast.

Family law. A patriarchal outlook is the basis of the traditional Islāmic law of family relationships. Fathers have the right to contract their daughters, whether minor or adult, in compulsory marriage. Only when a woman has been married before is her consent to her marriage necessary; but even then the father, or other marriage guardian, must conclude the contract on her behalf. In Ḥanafī and Shī'ī law, however, only minor girls may be contracted in compulsory marriage, and adult women may conclude their own marriage contracts, except that the guardian may have the marriage annulled if his ward has married beneath her social status.

Husbands have the right of polygamy and may be validly married at the same time to a maximum of four wives. Upon marriage a husband is obliged to pay to his wife her dower, the amount of which may be fixed by agreement or by custom; and during the marriage he is bound to maintain and support her provided she is obedient to him, not only in domestic matters but also in her general social activities and conduct. A wife who rejects her husband's dominion by leaving the family home without just cause forfeits her right to maintenance.

Law of divorce

But it is in the traditional law of divorce that the scales are most heavily weighted against the wife. A divorce may be effected simply by the mutual agreement of the spouses, which is known as *khul* when the wife pays some financial consideration to the husband for her release; and according to all schools except the Ḥanafīs a wife may obtain a judicial decree of divorce on the ground of some matrimonial offense—*e.g.*, cruelty, desertion, failure to maintain—committed by the husband. But the husband alone has the power unilaterally to terminate the marriage by repudiation (*ṭalāq*) of his wife. *Ṭalāq* is an extrajudicial process: a husband may repudiate his wife at will and his motive in doing so is not subject to scrutiny by the court or any other official body. A repudiation repeated three times constitutes a final and irrevocable dissolution of the marriage; but a single pronouncement may be revoked at will by the husband during the period known as the wife's *'iddah*, which lasts for three months following the repudiation (or any other type of divorce) or, where the wife is pregnant, until the birth of the child.

The legal position of children within the family group, as regards their guardianship, maintenance, and rights of succession, depends upon their legitimacy, and a child is legitimate only if it is conceived during the lawful wedlock of its parents. In Sunnī law no legal relationship exists between a father and his illegitimate child; but there is a legal tie, for all purposes, between a mother and her illegitimate child. Guardianship of the person (*e.g.*, control of education and marriage) and of the property of minor children belongs to the father or other close male, agnate relative, but the bare right of custody (*ḥaḍānah*) of young children, whose parents are divorced or separated, belongs to the mother or the female, maternal relatives.

Succession law. An individual's power of testamentary disposition is basically limited to one-third of his net estate (*i.e.*, the assets remaining after the payment of funeral expenses and debts) and two-thirds of the estate passes to the legal heirs of the deceased under the compulsory rules of inheritance.

Sunnī inheritance laws

There is a fundamental divergence between the Sunnī and the Shī'ī schemes of inheritance. Sunnī law is essentially a system of inheritance by male agnate relatives or *'aṣabah*—*i.e.*, relatives who, if they are more than one degree removed from the deceased, trace their connection with him through male links. Among the *'aṣabah*, priority is determined by: (1) class, descendants excluding ascendants, who in turn exclude brothers and their issue, who in turn exclude uncles and their issue; (2) degree, within each class the relative nearer in degree to the deceased excluding the more remote; (3) strength of blood tie, the germane, or full blood, connection excluding the half blood, or consanguine, connection among collateral relatives. This agnatic system is mitigated by allowing the surviving spouse and a limited number of females and nonagnates—the daughter; son's daughter; mother; grandmother; germane, consanguine, and uterine sisters; and uterine brother—to inherit a fixed fractional portion of the estate in suitable circumstances. But the females among these relatives only take half the share of the male relative of the same class, degree, and blood tie, and none of them excludes from inheritance any male agnate, however remote. No other female or non-agnatic relative has any right of inheritance in the presence of a male agnate. Where, for example, the deceased is survived by his wife, his daughter's son, and a distant agnatic cousin, the wife will be restricted to one-fourth of the inheritance, the grandson will be excluded altogether, and the cousin will inherit three-fourths of the estate.

Shī'ī law rejects the criterion of the agnatic tie and regards both the maternal and paternal connections as equally strong grounds of inheritance. In the Shī'ī system the surviving spouse always inherits a fixed portion, as in Sunnī law, but all other relatives, including females and nonagnates, are divided into three classes: (1) parents and lineal descendants; (2) grandparents, brothers and sisters, and their issue; (3) uncles and aunts and their issue. Any relative of class one excludes any relative of class two, who in turn excludes any relative of class three. Within each class the nearer in degree excludes the more remote, and the full blood excludes the half blood. While, therefore, a male relative normally takes double the share of the corresponding female relative, females and nonagnates are much more favourably treated than they are in Sunnī law. In the case mentioned above, for example, the wife would take one-fourth, but the remaining three-fourths would go to the daughter's son, or indeed to a daughter's daughter, and not to the agnatic cousin.

Under Shīʿī law the only restriction upon testamentary power is the one-third rule, but Sunnī law goes further and does not allow any bequest in favour of a legal heir. Under both systems, however, bequests that infringe these rules are not necessarily void and ineffective; the testator has acted beyond his powers, but the bequest may be ratified by his legal heirs.

Further protection is afforded to the rights of the legal heirs by the doctrine of death sickness. Any gifts made by a dying person in contemplation of his death are subject to precisely the same limitations as bequests, and, if they exceed these limits, will be effective only with the consent of the legal heirs.

Procedure and evidence. Traditionally, Sharīʿah law was administered by the court of a single *qāḍī*, who was the judge of the facts as well as the law, although on difficult legal issues he might seek the advice of a professional jurist, or *muftī*. There was no hierarchy of courts and no organized system of appeals. Through his clerk (*kātib*) the *qāḍī* controlled his court procedure, which was normally characterized by a lack of ceremony or sophistication. Legal representation was not unknown, but the parties would usually appear in person and address their pleas orally to the *qāḍī*.

Burden of proof

The first task of the *qāḍī* was to decide which party bore the burden of proof. This was not necessarily the party who brought the suit, but was the party whose contention was contrary to the initial legal presumption attaching to the case. In the case of an alleged criminal offense, for example, the presumption is the innocence of the accused, and in a suit for debt the presumption is that the alleged debtor is free from debt. Hence the burden of proof would rest upon the prosecution in the first case and upon the claiming creditor in the second. This burden of proof might, of course, shift between the parties several times in the course of the same suit, as, for example, where an alleged debtor pleads a counterclaim against the creditor.

The standard of proof required, whether on an initial, intermediate or final issue, was a rigid one and basically the same in both criminal and civil cases. Failing a confession or admission by the defendant, the plaintiff or prosecutor was required to produce two witnesses to testify orally to their direct knowledge of the truth of his contention. Written evidence and circumstantial evidence, even of the most compelling kind, were normally inadmissible. Moreover, the oral testimony (*shahādah*) had usually to be given by two male, adult Muslims of established integrity or character. In certain cases, however, the testimony of women was acceptable (two women being required in place of one man), and in most claims of property the plaintiff could satisfy the burden of proof by one witness and his own solemn oath as to the truth of his claim.

If the plaintiff or prosecutor produced the required degree of proof, judgment would be given in his favour. If he failed to produce any substantial evidence at all, judgment would be given for the defendant. If he produced some evidence, but the evidence did not fulfill the strict requirements of *shahādah*, the defendant would be offered the oath of denial. Properly sworn this oath would secure judgment in his favour; but if he refused it, judgment would be given for the plaintiff, provided, in some cases, that the latter himself would swear an oath.

In sum, the traditional system of procedure was largely self-operating. After his initial decision as to the incidence of the burden of proof, the *qāḍī* merely presided over the predetermined process of the law: witnesses were or were not produced, the oath was or was not administered and sworn, and the verdict followed automatically.

LAW IN CONTEMPORARY ISLĀM

The scope of Sharīʿah law and the mode of its administration. During the 19th century the impact of Western civilization upon Muslim society brought about radical changes in the fields of civil and commercial transactions and criminal law. In these matters the Sharīʿah courts were felt to be wholly out of touch with the needs of the time, not only because of their system of procedure and evidence but also because of the substance of the Sharīʿah doctrine, which they were bound to apply.

As a result, the criminal and general civil law of the Sharīʿah was abandoned in most Muslim countries and replaced by new codes based upon European models with a new system of secular tribunals to apply them. Thus, with the notable exception of the Arabian peninsula, where the Sharīʿah is still formally applied in its entirety, the application of Sharīʿah law in Islām has been broadly confined, from the beginning of the 20th century, to family law, including the law of succession at death and the particular institution of *waqf* endowments.

The decline of Sharīʿah

Nor, even within this circumscribed sphere, is Sharīʿah law today applied in the traditional manner. Throughout the Middle East generally Sharīʿah family law is now expressed in the form of modern codes, and it is only in the absence of a specific relevant provision of the code that recourse is had to the traditionally authoritative legal manuals. In India and Pakistan much of the family law is now embodied in statutory legislation, and since the law is there administered as a case-law system, the authority of judicial decisions has superseded that of the legal manuals.

In most countries, too, the court system has been, or is being, reorganized to include, for instance, the provision of appellate jurisdictions. In Egypt and Tunisia the Sharīʿah courts, as a separate entity, have been abolished, and Sharīʿah law is now administered through a unified system of national courts. In India, and, since partition, in Pakistan it has always been the case that Sharīʿah law has been applied by the same courts that apply the general civil and criminal law.

Finally, in many countries, special codes have been enacted to regulate the procedure and evidence of the courts that today apply Sharīʿah law. In the Middle East documentary and circumstantial evidence are now generally admissible; witnesses are put on oath and may be cross-examined, and the traditional rule that evidence is only brought by one side and that the other side, in suitable circumstances, takes the oath of denial has largely broken down. In sum, the court has a much wider discretion in assessing the weight of the evidence than it had under the traditional system of evidence. In India and Pakistan the courts apply the same rules of evidence to cases of Islāmic law as they do to civil cases generally. The system is basically English law, codified in the Indian Evidence Act, 1872.

Reform of Sharīʿah law. Traditional Islāmic family law reflected to a large extent the patriarchal scheme of Arabian tribal society in the early centuries of Islām. Not unnaturally certain institutions and standards of that law were felt to be out of line with the circumstances of Muslim society in the 20th century, particularly in urban areas where tribal ties had disintegrated and movements for the emancipation of women had arisen. At first this situation seemed to create the same apparent impasse between the changing circumstances of modern life and an allegedly immutable law that had caused the adoption of Western codes in civil and criminal matters. Hence, the only solution that seemed possible to Turkey in 1926 was the total abandonment of the Sharīʿah and the adoption of Swiss family law in its place. No other Muslim country, however, has as yet followed this example. Instead, traditional Sharīʿah law has been adapted in a variety of ways to meet present social needs.

From the outset the dominating issue in the Middle East has been the question of the juristic basis of reforms—*i.e.*, granted their social desirability, their justification in terms of Islāmic jurisprudential theory, so that the reforms appear as a new, but legitimate, version of the Sharīʿah.

In the early stages of the reform movement, the doctrine of *taqlīd* (unquestioning acceptance) was still formally observed and the juristic basis of reform lay in the doctrine of *siyāsah*, or "government," which allows the political authority (who, of course, has no legislative power in the real sense of the term) to make administrative regulations of two principal types.

The first type concerns procedure and evidence and restricts the jurisdiction of the Sharīʿah courts in the sense that they are instructed not to entertain cases that do not fulfill defined evidential requirements. Thus, an Egyptian law was enacted in 1931 that no disputed claim of marriage was to be entertained where the marriage could not be proved by an official certificate of registration, and no such certificate could be issued if the bride was less than 16 or the bridegroom less than 18 years of age at the time of the contract. Accordingly the marriage of a minor contracted by the guardian was still perfectly valid but would not, if disputed, be the subject of judicial relief from the courts. In theory the doctrine of the traditional authorities was not contradicted, but in practice an attempt had been made to abolish the institution of child marriage. The second type of administrative regulation was a directive to the courts as to which particular rule among existing variants they were to apply. This directive allowed the political authority to choose from the views of the different schools and jurists the opinion that was deemed best suited to present social circumstances. For example, the traditional Ḥanafī law in force in Egypt did not allow a wife to petition for divorce on the ground of any matrimonial offense committed by the husband, a situation that caused great hardship to abandoned or ill-treated wives. Mālikī law, however, recognizes the wife's right to judicial dissolution of her marriage on grounds such as the husband's cruelty, failure to provide maintenance and support, and desertion. Accordingly, an Egyptian law of 1920 codified the Mālikī law as the law henceforth to be applied by the Sharīʿah courts.

Egyptian reforms

By way of comparison, reform in the matters of child marriage and divorce was effected in the Indian subcontinent by statutory enactments that directly superseded the traditional Ḥanafī law. The Child Marriage Restraint Act, 1929, prohibited the marriage of girls below the age of 14 and boys below the age of 16 under pain of penalties; while the Dissolution of Muslim Marriages Act, 1939, modelled on the English Matrimonial Causes Acts, allowed a Ḥanafī wife to obtain judicial divorce on the standard grounds of cruelty, desertion, failure to maintain, etc.

In the Middle East, by the 1950s, the potential for legal reform under the principle of *siyāsah* had been exhausted. Since that time the basic doctrine of *taqlīd* has been challenged to an ever-increasing degree. On many points the law recorded in the medieval manuals, insofar as it represents the interpretations placed by the early jurists upon the Qurʾān and the *sunnah*, has been held no longer to have a paramount and exclusive authority. Contemporary jurisprudence has claimed the right to renounce those interpretations and to interpret for itself, independently and afresh in the light of modern social circumstances, the original texts of divine revelation: in short to reopen the door of *ijtihād* that had been in theory closed since the 10th century.

Syrian and Tunisian reforms

The developing use of *ijtihād* as a means of legal reform may be seen through a comparison of the terms of the Syrian law of Personal Status (1953) with those of the Tunisian Law of Personal Status (1957) in relation to the two subjects of polygamy and divorce by repudiation (*ṭalāq*).

As regards polygamy the Syrian reformers argued that the Qurʾān itself urges husbands not to take additional wives unless they are financially able to make proper provision for their maintenance and support. Classical jurists had construed this verse as a moral exhortation binding only on the husband's conscience. But the Syrian reformers maintained that it should be regarded as a positive legal condition precedent to the exercise of polygamy and enforced as such by the courts. This novel interpretation was then coupled with a normal administrative regulation that required the due registration of marriages after the permission of the court to marry had been obtained. The Syrian Law accordingly enacts: "The *qāḍī* may withhold permission for a man who is already married to marry a second wife, where it is established that he is not in a position to support them both." Far more extreme, however, is the approach of the Tunisian reformers. They argued that, in addition to a husband's financial ability to support a plurality of wives, the Qurʾān also required that co-wives should be treated with complete impartiality. This Qurʾānic injunction should also be construed, not simply as a moral exhortation, but as a legal condition precedent to polygamy, in the sense that no second marriage should be permissible unless and until adequate evidence was forthcoming that the wives would in fact be treated impartially. But under modern social and economic conditions such impartial treatment was a practical impossibility. And since the essential condition for polygamy could not be fulfilled the Tunisian Law briefly declares: "Polygamy is prohibited."

With regard to *ṭalāq* the Syrian law provided that a wife who had been repudiated without just cause might be awarded compensation by the court from her former husband to the maximum extent of one year's maintenance. The reform was once again represented as giving practical effect to certain Qurʾānic verses that had been generally regarded by traditional jurisprudence as moral rather than legally enforceable injunctions—namely, those verses that enjoin husbands to "make a fair provision" for repudiated wives and to "retain wives with kindness or release them with consideration." The effect of the Syrian law, then, is to subject the husband's motive for repudiation to the scrutiny of the court and to penalize him, albeit to a limited extent, for abuse of his power. Once again, however, the Tunisian *ijtihād* concerning repudiation is far more radical. Here the reformers argued that the Qurʾān orders the appointment of arbitrators in the event of discord between husband and wife. Clearly a pronouncement of repudiation by a husband indicated a state of discord between the spouses. Equally clearly the official courts were best suited to undertake the function of arbitration that then becomes necessary according to the Qurʾān. It is on this broad ground that the Tunisian law abolishes the right of a husband to repudiate his wife extrajudicially and enacts that: "Divorce outside a court of law is without legal effect." Although the court must dissolve the marriage if the husband persists in his repudiation, it has an unlimited power to grant the wife compensation for any damage she has sustained from the divorce—although in practice this power has so far been used most sparingly. In regard to polygamy and *ṭalāq* therefore, Tunisia has achieved by reinterpretation of the Qurʾān reforms hardly less radical than those effected in Turkey some 30 years previously by the adoption of the Swiss Civil Code.

Pakistani reforms

In Pakistan a new interpretation of the Qurʾān and *sunnah* was the declared basis of the reforms introduced by the Muslim Family Laws Ordinance of 1961, although the provisions of the Ordinance in relation to polygamy and *ṭalāq* are much less radical than the corresponding Middle Eastern reforms, since a second marriage is simply made dependent upon the consent of an Arbitration Council and the effect of a husband's repudiation is merely suspended for a period of three months to afford opportunity for reconciliation.

Judicial decisions in Pakistan have also unequivocally endorsed the right of independent interpretation of the Qurʾān. For example, in *Khurshīd Bībī* v. *Muḥammad Amīn* (1967) the Supreme Court held that a Muslim wife could as a right obtain a divorce simply by payment of suitable compensation to her husband. This decision was based on the Court's interpretation of a relevant Qurʾānic verse. But under traditional Sharīʿah law this form of divorce, known as *khulʿ*, whereby a wife pays for her release, is a contract between the spouses and as such entirely dependent upon the husband's free consent.

These are but a few examples of the many far-reaching changes that have been effected in the Islāmic family law. But the whole process of legal reform as it has so far developed still involves great problems of principle and practice. A hard core of traditionalist opinion still adamantly rejects the validity of the process of reinterpretation of the basic texts of divine revelation. The traditionalists argue that the texts are merely being manipulated to yield the meaning that suits the preconceived purposes of the reformers, and that therefore, contrary to fundamental Islāmic ideology, it is social desirability and not the will of Allāh that is the ultimate determinant of the law.

As regards the practical effect of legal reform, there exists in many Muslim countries a deep social gulf between a Westernized and modernist minority and the conservative mass of the population. Reforms that aim at satisfying the standards of progressive urban society have little significance for the traditionalist communities of rural areas or for the Mulsim fundamentalists, whose geographical and social distribution crosses all apparent boundaries. It is also often the case that the *qāḍīs*, through their background and training, are not wholly sympathetic with the purposes of the modernist legislators—an attitude often reflected in their interpretations of the new codes.

Such problems are, of course, inevitable in the transitional stage of social evolution in which Islām finds itself. But the one supreme achievement of jurisprudence over the past few decades has been the emergence of a functional approach to the question of the role of law in society. Jurisprudence has discarded the introspective and idealistic attitude that the doctrine of *taqlīd* had imposed upon it since early medieval times and now sees its task to be the solution of the problems of contemporary society. It has emerged from a protracted period of stagnation to adopt again the attitude of the earliest Muslim jurists, whose aim was to relate the dictates of the divine will to their own social environment. It is this attitude alone that has ensured the survival of the Sharīʿah in modern times as a practical system of law and that alone provides its inspiration for the future. (N.J.C.)

Social and ethical principles

FAMILY LIFE

A basic social teaching of Islām is the encouragement of marriage, and the Qurʾān regards celibacy definitely as something exceptional—to be resorted to only under economic stringency. Thus, monasticism as a way of life was severely criticized by the Qurʾān. With the appearance of Ṣūfism, however, many Ṣūfis preferred celibacy, and some even regarded women as an evil distraction from piety, although marriage remained the normal practice also with Ṣūfīs.

Pre-Islāmic practice of polygamy and the Qurʾān

Polygamy, which was practiced in pre-Islāmic Arabia, was permitted by the Qurʾān, which, however, limited the number of simultaneous wives to four, and this permission was made dependent upon the condition that justice be done among co-wives. The Qurʾān even suggests that "You shall never be able to do justice among women, no matter how much you desire." Medieval law and society, however, regarded this "justice" to be primarily a private matter between a husband and his wives, although the law did provide redress in cases of gross neglect of a wife. Right of divorce was also vested basically in the husband, who could unilaterally repudiate his wife, although the woman could also sue her husband for divorce before a court on certain grounds.

The virtue of chastity is regarded as of prime importance by Islām. The Qurʾān advanced its universal recommendation of marriage as a means to ensure a state of chastity (*iḥṣān*), which is held to be induced by a single free wife. The Qurʾān states that those guilty of adultery are to be severely punished with 100 lashes. Tradition has intensified this injunction and has prescribed this punishment for unmarried persons, but married adulterers are to be stoned to death. A false accusation of adultery is punishable by 80 lashes.

The general ethic of the Qurʾān considers the marital bond to rest on "mutual love and mercy," and the spouses are said to be "each other's garments." The detailed laws of inheritance prescribed by the Qurʾān also tend to confirm the idea of a central family—husband, wife, and children, along with the husband's parents. Easy access to polygamy (although the normal practice in Islāmic society has always been that of monogamy) and easy divorce on the part of the husband led, however, to frequent abuses in the family. In recent times, most Muslim countries have enacted legislation to tighten up marital relationships.

Rights of parents in terms of good treatment are stressed in Islām, and the Qurʾān extols filial piety, particularly tenderness to the mother, as an important virtue. A murderer of his father is automatically disinherited. The tendency of the Islāmic ethic to strengthen the immediate family on the one hand and the community on the other at the expense of the extended family or tribe did not succeed, however. Muslim society, until the encroachments upon it of modernizing influences, has remained basically one composed of tribes or quasi-tribes. Despite urbanization, tribal affiliations offer the greatest resistance to change and development of a modern polity. So strong, indeed, has been the tribal ethos that, in most Muslim societies, daughters are not given their inheritance share prescribed by the sacred law in order to prevent disintegration of the joint family's patrimony.

THE STATE

Because Islām draws no distinction between the religious and the temporal spheres of life, the Muslim state is by definition religious. The main differences between the Sunnī, Khawārij, and Shīʿī concepts of rulership have already been pointed out above. It should be noted that, although the office of the Sunnī caliph (*khalīfah*, one who is successor to the Prophet in rulership) is religious, this does not imply any functions comparable to those of the pope. The caliph has no authority either to define dogma or, indeed, even to legislate. He is the chief executive of a religious community, and his primary function is to implement the sacred law and work in the general interests of the community. He himself is not above the law and if necessary can even be deposed, at least in theory.

Sunnī political theory is essentially a product of circumstance—an after-the-fact rationalization of historical developments. Thus, between the Shīʿah legitimism that restricts rule to ʿAlī's family and the Khawārij democratism that allowed rulership to anyone, even to "an Ethiopian slave," Sunnism held the position that "rule belonged to the Quraysh" (the Prophet's tribe)—the condition that actually existed. Again, in view of the extremes represented by the Khawārij, who demanded rebellion against what they considered to be unjust or impious rule, and Shīʿites, who raised the *imām* to a metaphysical plane of infallibility, Sunnites took the position that a ruler has to satisfy certain qualifications but that rule cannot be upset on small issues. Indeed, under the impact of civil wars started by the Khawārij, Sunnism drifted to more and more conformism and actual toleration of injustice.

Political conformism of Sunnism

The first step taken in this direction by the Sunnites was the enunciation that "one day of lawlessness is worse than 30 years of tyranny." This was followed by the principle that "Muslims must obey even a tyrannical ruler." Soon, however, the sultan (ruler) was declared to be "shadow of God on earth." No doubt, the principle was also adopted—and insisted upon—that "there can be no obedience to the ruler in disobedience of God"; but there is no denying the fact that the Sunnī doctrine came more and more to be heavily weighted on the side of political conformism. This change is also reflected in the principles of legitimacy. Whereas early Islām had confirmed the pre-Islāmic democratic Arab principle of rule by consultation (*shūrā*) and some form of democratic election of the leader, those practices soon gave way to dynastic rule with the advent of the Umayyads. The *shūrā* was not developed into any institutionalized form and was, indeed, soon discarded. Soon the principle of "might is right" came into being, and later theorists frankly acknowledged that actual possession of effective power is one method of the legitimization of power.

In spite of this development, the ruler could not become absolute because a basic restraint was placed upon him by the Sharīʿah law under which he held his authority and which he dutifully was bound to execute and defend. When, in the latter half of the 16th century, the Mughal emperor Akbar in India wanted to arrogate to himself the right of administrative–legal absolutism, the strong reaction of the orthodox thwarted his attempt. In general, the *'ulamā'* (religious scholars) jealously upheld the sovereign position of the Sharī'ah against the political authority.

The effective shift of power from the caliph to the sultan was, again, reflected in the redefinition of the functions of the caliph. It was conceded that, if the caliph administered through *wazīrs* (viziers or ministers) or subordinate rulers (*amīrs*), it was not necessary for him to embody all the physical, moral, and intellectual virtues theoretically insisted upon earlier. In practice, however, the caliph was no more than a titular head from the middle of the 10th century onward, when real power passed to self-made and adventurous *amīrs* and sultans, who merely used the caliph's name for legitimacy.

EDUCATION

Muslim educational activity began in the 8th century, primarily in order to disseminate the teaching of the Qur'ān and the *sunnah* of the Prophet. The first task in this endeavor was to record the oral traditions and collect the written manuscripts. This information was systematically organized in the 2nd century AH, and in the following century a sound corpus was agreed upon. This vast activity of "seeking knowledge" (*ṭalab al-'ilm*) resulted in the creation of specifically Arab sciences of tradition, history, and literature.

When the introduction of the Greek sciences—philosophy, medicine, and mathematics—created a formidable body of lay knowledge, a creative reaction on the traditional religious base resulted in the rationalist theological movement of the Mu'tazilah. Based on that Greek legacy, from the 9th to the 12th century AD a brilliant philosophical movement flowered and presented a challenge to orthodoxy on the issues of the eternity of the world, the doctrine of revelation, and the status of the Sharī'ah.

Distinction between religious and secular sciences

The orthodox met the challenges positively by formulating the religious dogma. At the same time, however, for fear of heresies, they began to draw a sharp distinction between religious and secular sciences. The custodians of the Sharī'ah developed an unsympathetic attitude toward the secular disciplines and excluded them from the curriculum of the *madrasah* (college) system.

Their exclusion from the Sunnī system of education proved fatal, not only for those disciplines but, in the long run, for religious thought in general because of the lack of intellectual challenge and stimulation. A typical *madrasah* curriculum included logic (which was considered necessary as an "instrumental" science for the formal correctness of thinking procedure), Arabic literature, law, Ḥadīth, Qur'ān commentary, and theology. Despite sporadic criticism from certain quarters, the *madrasah* system remained impervious to change.

One important feature of Muslim education was that primary education (which consisted of Qur'ān reading, writing, and rudimentary arithmetic) did not feed candidates to institutions of higher education, and the two remained separate. In higher education, emphasis was on books rather than on subjects and on commentaries rather than on original works. This, coupled with the habit of learning by rote (which was developed from the basically traditional character of knowledge that encouraged learning more than thinking), impoverished intellectual creativity still further.

Despite these grave shortcomings, however, the *madrasah* produced one important advantage. Through the uniformity of its religio-legal content, it gave the *'ulamā'* the opportunity to effect that overall cohesiveness and unity of thought and purpose that, despite great variations in local Muslim cultures, has become a palpable feature of the world Muslim community. This uniformity has withstood even the serious tension created against the seats of formal learning by Ṣūfism through its peculiar discipline and its own centres.

In contrast to the Sunnī attitude toward it, philosophy continued to be seriously cultivated among the Shī'ah, even though it developed a strong religious character. Indeed, philosophy has enjoyed an unbroken tradition in Persia down to the present and has produced some highly original thinkers. Both the Sunnī and the Shī'ah medieval systems of learning, however, have come face to face with the greatest challenge of all—the impact of modern education and thought.

Organization of education developed naturally in the course of time. Evidence exists of small schools already established in the first century of Islām that were devoted to reading, writing, and instruction in the Qur'ān. These schools of "primary" education were called *kuttābs*. The well-known governor of Iraq at the beginning of the 8th century, the ruthless al-Ḥajjāj, had been a schoolteacher in his early career. When higher learning in the form of tradition grew in the 8th and 9th centuries, it was centred around learned men to whom students travelled from far and near and from whom they obtained a certificate (*ijāzah*) to teach what they had learned. Through the munificence of rulers and princes, large private and public libraries were built, and schools and colleges arose. In the early 9th century a significant incentive to learning came from the translations made of scientific and philosophical works from the Greek (and partly Sanskrit) at the famous *bayt al-ḥikmah* ("house of wisdom") at Baghdad, which was officially sponsored by the caliph al-Ma'mūn. The Fāṭimid caliph al-Ḥākim set up a *dār alḥikmah* ("hall of wisdom") in Cairo in the 10th–11th centuries. With the advent of the Seljuq Turks, the famous vizier Niẓām al-Mulk created an important college at Baghdad, devoted to Sunnī learning, in the latter half of the 11th century. One of the world's oldest surviving universities, al-Azhar at Cairo, was originally established by the Fāṭimids, but Saladin (Ṣalāḥ ad-Dīn al-Ayyūbī), after ousting the Fāṭimids, consecrated it to Sunnī learning in the 12th century. Throughout subsequent centuries, colleges and quasi-universities (called *madrasah* or *dār al-'ulūm*) arose throughout the Muslim world from Spain (whence philosophy and science were transmitted to the Latin West) across Central Asia to India.

In Turkey a new style of *madrasah* came into existence; it had four wings, for the teaching of the four schools of Sunnī law. Professorial chairs were endowed in large colleges by princes and governments, and residential students were supported by college endowment funds. A myriad of smaller centres of learning were endowed by private donations.

CULTURAL DIVERSITY

Underneath the legal and creedal unity, the world of Islām harbours a tremendous diversity of cultures, particularly in the outlying regions. The expansion of Islām can be divided into two broad periods. In the first period of the Arab conquests, the assimilative activity of the conquering religion was far-reaching. Although Persia resurrected its own language and a measure of its national culture after the first three centuries of Islām, its culture and language had come under heavy Arab influence. Only after Ṣafavid rule installed Shī'ism as a distinctive creed in the 16th century did Persia regain a kind of religious autonomy. The language of religion and thought, however, continued to be Arabic.

Ṣūfī compromises with local customs

In the second period, the spread of Islām was not conducted by the state with *'ulamā'* influence but was largely the work of Ṣūfī missionaries. The Ṣūfīs, because of their latitudinarianism, compromised with local customs and beliefs and left a great deal of the pre-Islāmic legacy in every region intact. Thus, among the Central Asian Turks, shamanistic practices were absorbed, while in Africa the holy man and his *barakah* (an influence supposedly causing material and spiritual well-being) are survivors from the older cults. In India there are large areas geographically distant from the Muslim religio-political centre of power in which customs are still Hindu and even pre-Hindu and in which people worship a motley of saints and deities in common with the Hindus. The custom of *satī*, under which a widow burned herself alive along with her dead husband, persisted in India even among some Muslims until late into the Mughal period. The 18th- and 19th-century reform movements exerted themselves to "purify" Islām of these accretions and superstitions.

Indonesia affords a striking example of this phenomenon. Because Islām reached there late and soon thereafter came under European colonialism, the Indonesian society has retained its pre-Islāmic world view beneath an overlay of Islāmic practices. It keeps its customary law (called *adat*) at the expense of the Sharī'ah; many of its tribes are still matriarchal; and culturally the Hindu epics *Rāmāyana* and *Mahābhārata* hold a high position in national life. Since the 19th century, however, orthodox Islām has gained steadily in strength because of fresh contacts with the Middle East.

Apart from regional diversity, the main internal division within Islāmic society is brought about by urban and village life. Islām originally grew up in the two cities of Mecca and Medina, and as it expanded, its peculiar ethos appears to have developed in urban areas. Culturally, it came under a heavy Persian influence in Iraq, where the Arabs learned the ways and style of life of their conquered people, who were culturally superior to them. The custom of veiling women (which originally arose as a sign of aristocracy but later served the purpose of segregating women from men—the *pardah*), for example, was acquired in Iraq.

Another social trait derived from outside cultures was the disdain for agriculture and manual labour in general. Because the people of the town of Medina were mainly agriculturists, this disdain could not have been initially present. In general, Islām came to appropriate a strong feudal ethic from the peoples it conquered. Also, because the Muslims generally represented the administrative and military aristocracy and because the learned class (the *'ulamā'*) was an essential arm of the state, the higher culture of Islām became urban based.

This city orientation explains and also underlines the traditional cleavage between the orthodox Islām of the *'ulamā'* and the folk Islām espoused by the Ṣūfī orders of the countryside. In the modern period, the advent of education and rapid industrialization threatened to make this cleavage still wider. With the rise of a strong and widespread fundamentalist movement in the second half of the 20th century, this dichotomy has decreased.

Religion and the arts

THE VISUAL ARTS

The Arabs before Islām had hardly any art except poetry, which had been developed to full maturity and in which they took great pride. As with other forms of culture, the Muslim Arabs borrowed their art from Persia and Byzantium. Whatever elements the Arabs borrowed, however, they Islāmized in a manner that fused them into a homogeneous spiritual-aesthetic complex. The most important principle governing art was aniconism; *i.e.*, the religious prohibition of figurization and representation of living creatures. Underlying this prohibition is the assumption that God is the sole author of life and that a person who produces a likeness of a living being seeks to rival God. The tradition ascribed to the Prophet that a person who makes a picture of a living thing will be asked on the Day of Judgment to infuse life into it, whether historically genuine or not, doubtless represents the original attitude of Islām. In the Qur'ān (3:49, 5:113), reflecting an account in a New Testament apocryphal work, it is counted among the miracles of Jesus that he made likenesses of birds from clay "by God's order," and, when he breathed into them, they became real birds, again, "by God's order."

Aniconism as a governing principle of art

Hence, in Islāmic aniconism two considerations are fused together: (1) rejection of such images that might become idols (these may be images of anything) and (2) rejection of figures of living things. Plato and Plotinus, Greek philosophers, had also dismissed representative art as an "imitation of nature"; *i.e.*, as something removed from reality. The Islāmic attitude is more or less the same, with the added element of attributing to the artist a violation of the sanctity of the principle of life. The same explanation holds for the Qur'ānic criticism of a certain kind of poetry, namely, free indulgence in extravagant image mongering: "They [poets] recklessly wander in every valley" (26:225).

This basic principle has, however, undergone modifications. First, pictures were tolerated if they were confined to private apartments and harems of palaces. This was the case with some members of the Umayyad and 'Abbāsid dynasties, Turks, and Persians—in particular with the Shī'ah, who have produced an abundance of pictorial representations of the holy family and of the Prophet himself. Second, in the field of pictorial representation, animal and human figures are combined with other ornamental designs such as fillets and arabesques—stressing their ornamental nature rather than representative function. Third, for the same reason, in plastic art they appear in low relief. In other regions of the Muslim world—in North Africa, Egypt, and India (except for Mughal palaces)—representational art was strictly forbidden. Even in paintings, the figures have little representational value and are mostly decorative and sometimes symbolic. This explains why plastic art is one of the most limited areas of Islāmic art. The only fullfledged plastic figures are those of animals and a few human figures that the Seljuqs brought from eastern Turkistan.

Much more important than plastic art were paintings, particularly frescoes and later Persian and Perso-Indian miniatures. Frescoes are found in the Umayyad and 'Abbāsid palaces and in Spain, Iran, and in the harem quarters of the Mughal palaces in India. Miniature paintings, introduced in Persia, assumed much greater importance in the later period in Mughal India and Turkey. Miniature painting was closely associated with the art of book illumination, and this technique of decorating the pages of the books was patronized by princes and other patrons from the upper classes. (Miniature painting is also discussed below; see *Illustration of myth and legend.*)

Miniature painting

MUSIC

Instrumental music was forbidden by the orthodox in the formative stages of Islām. As for vocal music, its place was largely taken by a sophisticated and artistic form of the recitation of the Qur'ān known as *tajwīd.* Nevertheless, the Muslim princely courts generously patronized and cultivated music. Arab music was influenced by Persian and Greek music. Al-Fārābī, a 10th-century philosopher, is credited with having constructed a musical instrument called the *arghanūn* (organ). In India, Amīr Khosrow, a 14th-century poet and mystic, produced a synthesis of Indian and Persian music and influenced the development of later Indian music.

Among the religious circles, the Ṣūfīs introduced both vocal and instrumental music as part of their spiritual practices. The *samā',* as this music was called, was opposed by the orthodox at the beginning, but the Ṣūfīs persisted in this practice, which slowly won general recognition. The great Ṣūfī poet Jalāl ad-Dīn ar-Rūmī (died 1273)—revered equally by the orthodox and the Ṣūfīs—heard the divine voice in his stringed musical instrument when he said "Its head, its veins (strings) and its skin are all dry and dead; whence comes to me the voice of the Friend?"

LITERATURE

In literature, drama and pure fiction were not allowed—drama because it was a representational art and fiction because it was akin to lying. Similar constraints operated against the elaboration of mythology (see below *Islāmic myth and legend*). Story literature was tolerated, and the great story works of Indian origin—*The Thousand and One Nights* and *Kalīlah wa Dimnah*—were translated from the Persian, introducing secular prose into Arabic. Didactic and pious stories were used and even invented by popular preachers. Much of this folklore found its way back into enlarged editions of *The Thousand and One Nights* and, through it, has even influenced later history writing. Because of the ban on fictional literature, there grew a strong tendency in later literary compositions—in both poetry and prose—toward hyperbole (*mubālaghah*), a literary device to satisfy the need of getting away from what is starkly real without committing literal falsehood, thus often resulting in the caricature and the grotesque. Poetry lent itself particularly well to this device, which was freely used in panegyrics, satires, and lyrics. As a form of effective expression, poetry is eminently characteristic of the East. The Arab genius is almost natively poetical with its strong and vivid imagination not easily amenable to the rigorous order that reason imposes upon the mind.

Poetry as a characteristic of the East

This borderline attitude between the real and the unreal was particularly favourable to the development, in all medieval Islāmic literatures of the Middle East, of the lyric and panegyric forms of poetry wherein every line is a self-contained unit. Much more importantly, it afforded a specially suitable vehicle for a type of mystic poetry in which it is sometimes impossible to determine whether the poet is talking of earthly love or spiritual love. For the same reason, poetry proved an effective haven for thinly veiled deviations from and even attacks on the literalist religion of the orthodox.

ARCHITECTURE

Architecture is by far the most important expression of Islāmic art, particularly the architecture of mosques. It illustrates both the diversity of cultures that participated in the Islāmic civilization and the unifying force of Islāmic monotheism represented by the spacious expanse of the mosque—a veritable externalization of the all-enveloping divine unity, heightened by the sense of infinity of the arabesque design. The arabesque, though ornately decorative, spiritually represents the infinite vastness of God.

Among the earliest monuments are the mosque of ʿAmr built in Egypt in 641–642 and the famous Dome of the Rock of Jerusalem (finished in 691), which, however, is not a mosque but a monument, a concentric-circular structure consisting of a wooden dome set on a high drum and resting on four tiers and 12 columns. The Umayyad ruler al-Walīd (died 715) built the great mosque at Damascus and al-Aqṣā Mosque at Jerusalem with two tiers of arcades in order to heighten the ceiling. The early Syro-Egyptian mosque is a heavily columned structure with a prayer niche (*miḥrāb*) oriented toward the Kaʿbah sanctuary at Mecca.

In Spanish and North African architecture these features are combined with Roman-Byzantine characteristics, the masterpieces of Spanish architecture being the famous Alhambra Palace at Granada and the Great Mosque of Córdoba. In the famous Persian mosques, the characteristic Persian elements are the tapered brick pillars, the arches (each supported by several pillars), the huge arcades, and the four sides called *eyvān*s. With the advent of the Seljuqs in the 11th century, faience decoration (glazed earthenware) of an exquisite beauty was introduced, and it gained further prominence under the Timurids (14th–16th centuries).

In the number and greatness of mosques, Turkey has the pride of first place in the Muslim world. Turkey began with a Persian influence and then later Syrian in the 13th and 14th centuries, but Turkey developed its own cupola domes and monumental entrances. The Turkish architects accomplished symmetry by means of one large dome, four semidomes, and four small domes among them. In the Indo-Pakistan subcontinent, Muslim architecture first employed Hindu architectural features (*e.g.*, horizontal rather than arcuate, or bowlike, arches and Hindu ornamentation), but later the Persian style predominated.

(F.R./Ed.)

Islāmic myth and legend

The strict monotheism of Islām does not allow for much mythological embellishment, and only reluctantly were the scriptural revelations of the Qurʾān elaborated and enlarged by commentators and popular preachers. Thus, in the first three centuries, a number of ideas from the ancient Near East, from Hellenistic and especially from Judeo-Christian traditions were absorbed into Islām and given at least partial sanction by the theologians. At the same time, legends were woven around the Prophet Muḥammad and the members of his family. Though inconsistent with historical reality, these legends formed for the masses the main sources of inspiration about the famous figures of the past.

Role of storytellers

Since early times Islāmic theologians have sought to disregard the Qurʾānic interpretation of both storytellers and mystics. The *quṣṣāṣ*, or storytellers, made the Qurʾānic revelation more understandable to the masses by filling in the short texts with detailed descriptions that were not found in scripture. Though the mystics tried to maintain the purity of the divine word, they also attempted a spiritualization of both the Qurʾān and the popular legends that developed around it. Their way of giving to the Qurʾānic words a deeper meaning, however, and discovering layer after layer of meaning in them, sometimes led to new quasi-mythological forms. Later Islāmic mystical thinkers built up closed systems that can be called almost mythological (*e.g.*, the angelology—theory of angels—of Suhrawardī al-Maqtūl, executed 1191). An interesting development is visible in poetry, especially in the Persian-speaking areas, where mythological figures and pious legend often were turned into secular images that might awaken in the reader a reminiscence of their religious origin. Such images contribute to the iridescent and ambiguous character of Persian poetry.

SOURCES AND VARIATIONS

The Qurʾān and non-Islāmic influences. The sources of Islāmic mythology are first of all the Qurʾānic revelations. Since, for the Muslims, the Qurʾān is the uncreated word of God (the text revealed to Muḥammad considered an earthly manifestation of the eternal and uncreated original in heaven), it contains every truth, and whatever is said in it has been the object of meditation and explanation through the centuries. Thus, since the 9th century, commentators on the Qurʾān have been by far the most important witnesses for Islāmic "mythology." They wove into their explanations various strands of Persian and ancient oriental lore and relied heavily on Jewish tradition. For example, the Jewish convert, Kaʿb al-Aḥbār brought much of the *Isrāʾīliyāt* (things Jewish) into Islāmic tradition. Later on, the mystics' commentaries expressed some gnostic (a dualistic viewpoint in which spirit is viewed as good and matter as evil) and Hellenistic concepts, of which the Hellenistic idea of the Perfect Man—personified in Muḥammad—was to gain greatest prominence. Commentaries written in the border areas of Islāmic countries now and then accepted a few popular traditions from their respective areas; however, the formative period was finished quite early. Traditions about the life and sayings of the Prophet grew larger and larger and are interesting for the study of the adoption of foreign mythological material. A valuable source for Islāmic legends are the *qiṣaṣ al-anbiyāʾ*—stories of the prophets, such as those by Thaʿālibī (born 1035) and Kisāʾī (11th century)—traditions concerning the prophets of yore in which a large number of pre-Islāmic and non-Islāmic ideas were incorporated.

While the classical mythology of Islām, as far as it can be properly called so, is spread over the whole area of Islām, the miracles and legends around a particular Muslim saint are found chiefly in the area of his special influence (especially where his order is most popular). Even if the names of the saints differ, the legends woven around them are very similar to each other and almost interchangeable. In the area where Persian was read—from Ottoman Turkey to India—the mythological concepts of Ferdowsī's *Shāh-nāmeh* are found side by side with the legends taken from ʿAṭṭār's and Rūmī's works.

Trans-cultural prolifera-tion

The mystics. From the 11th century onward, the biographies of the mystics often show interesting migrations of legendary motifs from one culture to another. For the Persian-speaking countries the *Tazkerat ol-Owlīyāʾ* ("Memoirs of the Saints") of Farīd od-Dīn ʿAṭṭār (died *c.* 1220) has become the storehouse of legendary material about the early Ṣūfī mystics. ʿAṭṭār's Persian epics (especially his *Manṭeq ot-ṭeyr*, the "Birds' Conversation") also contain much material that was used by almost every writer after him. *The Maṡnavī* (a sort of poetic encyclopaedia of mystical thought in 26,000 couplets) of Jalāl od-Dīn Rūmī (died 1273) is another important source for legends of saints and prophets. For the Iranian world view, Ferdowsī's (died *c.* 1020) *Shāh-nāmeh* ("Book of Kings") gave a poetical account of the mythology of old Iran, and its heroes became models for many poets and writers. The whole mythological and legendary heritage is condensed in allusions found in lyrical and panegyrical poetry. The Persian poet Ebrāhim ebn ʿAlī Khāqānī's (*c.* 1121–*c.* 1199) works, *qaṣīdahs* ("Odes"), are typical. The close connection of the Ṣūfī orders with the artisans' lodges and guilds was instrumental in the dissemination of legendary material, especially about the alleged founder, or patron, of the guild (such as Ḥallāj as patron of cottoncarders and Idrīs as patron of the tailors).

Muslim historians interested in world history often began their works with mythological tales; central Asian traditions were added in Iran during the Il-Khanid period (AD 1256–1335). Folk poetry, in the different languages spoken by Muslims, provides a popular representation of traditional material, be it in Arabic, Persian, Turkish, the Indian and Pakistani languages (Urdu, Bengali, Sindhi, Panjabi, Baluchi, etc.), or the African languages; in all of them allusions to myth and legend are found down to the level of riddles and lullabies. Typical of the legendary tradition of the Shīʿah are the *taʿziyas* ("passion-plays") in Iran, commemorating the death of Husayn ibn ʿAlī in Karbalāʾ (680) and the *marsīyehs* (threnodies or elegies for the dead), which form an important branch of the Urdu poetry of India and Pakistan. A proper study of the distribution of most aspects of mythology in the various Muslim areas has not been undertaken, since much of the popular material is rarely available in print or is written in less-known languages—a good example is the extremely rich collections of legends and popular pious works in the Pakistani language, Sindhi.

TYPES OF MYTH AND LEGEND

Cosmogony and eschatology. The world was created by God's word *kun* ("Be") out of nothing; after the creation of the angelic beings from light, Adam was formed from clay and destined to be God's vicegerent, *khalīfah*. All of the angels obeyed God's order to prostrate themselves before Adam, except Iblīs (Satan), who refused and was cursed; due to Iblīs' instigation Adam ate the forbidden fruit (or grain) and was driven out of paradise. Questions of original sin or of Eve's role do not arise in the Muslim version of creation. Satan's disobedience has been explained by the mystics as actually an expression of his obedience to the divine will that does not allow worship of any but the Lord and that conflicted with the order that Satan prostrate himself before Adam.

By courtesy of the Bibliothèque Nationale, Paris

Satan's refusal to worship Adam, depicting the rebellious angel Iblīs, or Satan, as the human figure on a prayer rug (right). From a 17th-century manuscript of Majāles ol-ʿOshshāq. In the Bibliothèque Nationale, Paris (supplément Persan 1559).

Before the creation, God addressed the posterity of Adam: "Am I not your Lord," *alastu birabbikum,* and they answered "Yes" (Qurʾān, *sūrah* 7:172). This pre-eternal covenant is the favourite topic of mystical poetry, especially in the Persian-speaking areas for expressing pre-eternal love between God and man, or the unchangeable fate that was accepted that very day, the Yesterday as contrasted to the Tomorrow of resurrection. Angels and jinns (genies) are living powers that become visible in human life; they are accepted as fully real.

Every destiny is written on the "well-preserved tablet," and now "the pen has dried up"—a change in destiny is not possible. Later mystics have relied on an extra-Qurʾānic revelation in which God attests: "I was a hidden treasure" and have seen the reason for creation in God's yearning to be known and loved. For them, creation is the projection of divine names and qualities onto the world of matter.

Centrality of death and resurrection

The central event of Islām is death and resurrection. The dead will be questioned by two terrible angels (that is why the profession of faith is recited to the dying); only the souls of martyrs go straight to heaven where they remain in the crops of green birds around the divine throne (green is always connected with heavenly bliss). The end of the world will be announced by the coming of the mahdī (literally, "the directed or guided one")—a messianic figure who will appear in the last days and is not found in the Qurʾān but developed out of Shīʿah speculations and sometimes identified with Jesus. The mahdī will slay the Dajjāl, the one-eyed evil spirit, and combat the dangerous enemies, Yājūj and Mājūj, who will come from the north of the earth. The trumpet of Isrāfīl, one of the four archangels, will awaken the dead for the day of resurrection, which is many thousands of years long and the name of which has come to designate a state of complete confusion and turmoil.

The eschatological inventory as described in the Qurʾān was elaborated by the commentators: the scales on which the books or deeds are weighed (an old Egyptian idea), the book in which the two recording angels have noted down man's deeds, and the narrow bridge that is said to be sharper than a sword and thinner than a hair and leads over hell (an Iranian idea). The dreadful angels of hell and the horrors of that place are as thoroughly described by theologians as the pleasures of paradise, with its waters and gardens and the houris who are permanent virgins. Pious tradition promises space in heavenly mansions, filled with everything beautiful, to those who repeat certain prayer formulas a certain number of times, or for similar rewarding deeds, whereas the mystic longs not "for houris some thousand years old" but for the vision of God, who will be visible like the full moon. In the concept of the *sidrah* tree as the noblest place in paradise a remnant may be found of the old tree of life. God's throne is on the waters (Qurʾān, *sūrah* 11:9) in the highest world, surrounded by worshipping angels. The created world, the earth, is surrounded by the mountain Qāf and enclosed by two oceans that are separated by a barrier. Mecca is the navel of the earth, created 2,000 years before everything else, and the deluge did not reach to proto-*Kaʿbah*. Often the world is conceived as a succession of seven heavens and seven earths, and a popular tradition says that the earth is on water, on a rock, on the back of a bull, on a *kamkam* (meaning unknown), on a fish, on water, on wind, on the veil of darkness—hence the Persian expression *az māh tā māhī,* from the moon to the fish; *i.e.,* throughout the whole world.

Tales and legends concerning religious figures. The majority of popular legends concern the leading personalities of Islām.

Muhammad. Muḥammad, whose only miracle, according to his own words, was the bringing of the Qurʾān, is credited with innumerable miracles and associated with a variety of miraculous occurrences: his finger split the moon, the cooked poisoned meat warned him not to touch it, the palm trunk sighed, the gazelle spoke for him; he cast no shadow; from his perspiration the rose was created, etc. His ascension to heaven (*miʿrāj*) is still celebrated: he rode the winged horse Burāq in the company of the angel Gabriel through the seven spheres, meeting the other prophets there, until he reached the divine presence, alone, even without the angel of inspiration. Muḥammad-mysticism proper was developed in the late 9th century; he is shown as the one who precedes creation, his light is pre-eternal, and he is the reason for and goal of creation. He becomes the perfect man, uniting the divine and the human sphere as dawn is between night and day. His birth was surrounded by miracles, and his birthday (12. Rabīʿ I) became a popular holiday on which numerous poems were written to praise his achievements. The hope for him who has been sent as "mercy for the worlds" and will intercede for his community on Doomsday is extremely strong, especially among the masses, where these legends have completely overshadowed his historical figure.

Muḥammad-mysticism

Other Qurʾānic figures. In addition to Muḥammad himself, his cousin and son-in-law ʿAlī, the Shīʿah hero, has been surrounded by legends concerning his bravery, his miraculous sword, Dhūaʾl-fiqār, and his wisdom. ʿAlī's son, Ḥusayn, is the subject of innumerable poems that concern the day of his final fight in Karbalāʾ.

Almost every figure mentioned in the Qurʾān has become the centre of a circle of legends, be it Yūsuf, the symbol of overwhelming beauty, or Jesus with the lifegiving breath, the model of poverty and asceticism. Of special interest is Khiḍr, identified with the unnamed companion of Moses (Qurʾān, *sūrah* 20). He is the patron saint of the wayfarers, connected with green, the colour of heavenly bliss, appearing whenever a pious person is in need, and immortal since he drank from the fountain of life, which is hidden in the darkness. In many respects, he is the Islāmic counterpart of Elijah. Strong influences of the Alexander romances (a widely distributed literary genre dealing with the adventures of Alexander the Great) are visible in his figure.

Mystics and other later figures. The great religious personalities have become legendary, especially the martyr-mystic Ḥallāj (executed in Bagdad, 922). His word *anā al-Ḥaqq,* "I am the Creative Truth," became the motto of many later mystics. His death on the gallows is the model for the suffering of lovers, and allusions to his fate are frequent in Islāmic literature. An earlier mystic, Abū Yazīd al-Bisṭāmī (died 874), was the first to speak about the ascension of the mystic to heaven, which is a metaphor for higher unitive, mystical experience. A variation of the Buddha legend has been transferred onto the person of the first Ṣūfī (mystic) who practiced absolute poverty and trust in God, the Central Asian Ibrāhīm ibn Adham (died *c.* 780). The founders of mystical orders were credited by their followers with a variety of miracles, such as riding on lions, healing the sick, walking on water, being present at two places at the same time, and cardiognosia (which is the knowledge of what is in another's heart, or thought reading). ʿAbd al-Qādir al-Jīlānī (died 1166), the founder of the widespread Qādirīyah order of mystics, and many others have attracted upon themselves a large number of popular stories that formerly had been told about pre-Islāmic saints or about some divinities, and these motifs can easily be transferred from one person to the other. In this sphere the survival of pre-Islāmic customs and legends is most visible. The idea of the hierarchy of saints, culminating in the *quṭb*, the pole or axis, thanks to whose activities the world keeps going, belongs to the mythology of Ṣūfism (Islāmic mysticism).

The mystics as miracle workers

Mythologization of secular tales. A feature of Islāmic mythology is the transformation of unreligious stories into vehicles of religious experience. The old hero of romantic love in Arabic literature, Majnūn, "the demented one," became a symbol of the soul longing for identification with God, and in the Indus Valley the tales of Sassui or Sohnī, the girls who perish for their love, and other romantic figures, have been understood as symbols of the soul longing for union with God through suffering and death.

Tales and beliefs about numbers and letters. Many Muslim tales, legends, and traditional sayings are built upon the mystical value of numbers, such as the threefold or sevenfold repetition of a certain rite. This is largely explained by examples from the life of a saintly or pious person, often the Prophet himself, who used to repeat this or that formula so and so many times. The number 40, found in the Qurʾān (as also in the Bible) as the length of a period of repentance, suffering, preparation, and steadfastness, plays the same role in Islām where it is connected, for example, with the 40 days' preparation and meditation, or fasting, of the novice in the mystical brotherhood. To each number, as well as to each day of the week, special qualities are attributed through the authority of both actual and alleged statements of the Prophet. Many pre-Islāmic customs were thus justified. The importance given to the letters of the Arabic alphabet is peculiar to Muslim pious thought. Letters of the alphabet were assigned numerical values: the straight *alif* (numerical value one), the first letter of the alphabet, becomes a symbol of the uniqueness and unity of Allāh; the *b* (numerical value two), the first letter of the Qurʾān, represents to many mystics the creative power by which everything came into existence; the *h* (numerical value five) is the symbol of *huwa,* He, the formula for God's absolute transcendence; the *m* (numerical value 40) is the "shawl of humanity" by which God, the One (al-Aḥad), is separated from *Aḥmad* (*Muḥammad*). M is the letter of human nature and hints at the 40 degrees between man and God. The sect of the Ḥurūfīs developed these cabalistic interpretations of letters, but they are quite common in the whole Islāmic world and form almost a substitute for mythology.

Number symbolism

ILLUSTRATION OF MYTH AND LEGEND

Since the art of representation is opposed in Islām, illustrations of mythological and legendary subjects are rarely found. Miniature painting developed only in the Persian and, later on, in the Turkish and Indo-Muslim areas. Books such as Zakarīyāʾ ebn Moḥammad al-Qazvīnī's *Cosmography* contain in some manuscripts a few pictures of angels, like Isrāfīl with the trumpet, and histories of the world or histories of the prophets, written in Iran or Turkey, also contain in rare manuscripts representations of angels or of scenes as told in the Qurʾān, especially the story of Yūsuf and Zalīkhā, which inspired many poems. The *Shāh-nāmeh* has been fairly frequently illustrated. When the Prophet of Islām is shown at all, his face is usually covered and in several cases his companions or his family members are also shown with veiled faces.

Persian, Turkish, and Indian Mughal miniature painting

The only subject from the legends surrounding Muḥammad that has been treated by miniaturists several times is his ascension to heaven. There are a number of splendid Persian miniatures depicting this. In poetical manuscripts that contain allusions to legends of the saints, these topics were also sometimes illustrated (*e.g.,* Jonah and the great fish or scenes from the wanderings of Khiḍr). Several miniatures deal with the execution of the mystic al-Ḥallāj. Mythological themes proper are found almost exclusively in the paintings of Mughal India; especially in the period of Jahāngīr, in which the eschatological peace of lion and lamb lying together is illustrated as well as the myth of the earth resting on the bull, on the fish, etc. But by that time European influence was also already visible in Mughal art.

SIGNIFICANCE AND MODERN INTERPRETATIONS

Mythology proper has only a very small place in official Islām and is mostly an expression of popular traditions through which pre-Islāmic influences seeped into Islām. Reformers tried to purge Islām of all non-Qurʾānic ideas and picturesque elaborations of the texts, whereas the mystics tried to spiritualize them as far as possible. Modern Muslim exegesis attempts to interpret many of the mythological strands of the Qurʾān in the light of modern science, as psychological factors, like Muḥammad's ascension to heaven, and especially deprives the eschatological parts of the Qurʾān of their religious significance. Cosmic events are interpreted as predictions of modern scientific research. To some interpreters, jinns and angels are spiritual forces; to others, jinns are microbes or the like. Thus the religious text is confused with a textbook of science. Popular legends surrounding the Prophet and the saints are still found among the masses but are tending to disappear under the influence of historical research, though many of them have formed models for the behaviour and spiritual life of the Muslim believer. (An.Sc.)

BIBLIOGRAPHY

Islām. *General works:* *Cambridge History of Islam,* (1970, vol. 2, pt. 8; reissued 1977, vol. 2B), an excellent survey; MARSHALL G.S. HODGSON, *The Venture of Islam,* 3 vol. (1974), a major and influential study of the religion and civilization; R.M. SAVORY (ed.), *Introduction to Islamic Civilization* (1976), a collection of scholarly articles on Islāmic history, religion, literature, language, and other topics; BERNARD LEWIS (ed.), *The World of Islam: Faith, People, Culture* (U.S. title, *Islam and the Arab World,* 1976), a collection of articles on various aspects of Islāmic culture, and *Islam: From the Prophet Muhammad to the Capture of Constantinople,* 2 vol. (1974), a history composed of translations of original sources; W. MONTGOMERY WATT, *The Majesty That Was Islam: The Islamic World, 661–1100* (1974), a concise introductory history of the rise and decline of the Islāmic Empire; HAMILTON A.R. GIBB, *Mohammedanism,* 2nd ed. (1953, reissued with revisions 1969), a penetrating and concise account of the development of Islām; LOUIS GARDET, *Mohammedanism,* trans. by WILLIAM BURRIDGE (1961), a systematic presentation of Islām, with religious insight; FAZLUR RAHMAN, *Islam,* 2nd ed. (1979), a historical and systematic interpretation of Islām, and *Islamic Methodology in History* (1965), a critical appraisal of the development of *sunnah, ijmā‘,* and *ijtihād;* REUBIN LEVY, *An Introduction to the Sociology of Islam* (1930–), useful account of the development of Islāmic society and institutions. JOHN W. BAGNOLE, *Cultures of the Islamic Middle East* (1978), an annotated guide to 402 English-language readings for the nonspecialist.

Education: ARTHUR S. TRITTON, *Materials on Muslim Education in the Middle Ages* (1957), an informative, useful compilation; BAYARD DODGE, *Muslim Education in Medieval Times* (1962), a useful sketch.

Political theory and institutions: ERWIN I.J. ROSENTHAL, *Political Thought in Medieval Islam* (1958), a good general survey of the subject.

Islāmic arts: In view of the wealth of descriptive treatments, rather than theory, it is difficult to point to a single source. K.A.C. CRESWELL, *A Bibliography of the Architecture, Arts and Crafts of Islam to 1st Jan. 1960* (1961), and *Supplement, Jan. 1960 to Jan. 1972* (1973), contain all the necessary references. See also his *Early Muslim Architecture,* 2nd ed. (1969), and AMERICAN UNIVERSITY AT CAIRO, CENTER FOR ARABIC STUDIES, *Studies in Islamic Art and Architecture in Honor of Professor K.A.C. Creswell* (1965); HAMILTON A.R. GIBB, *Arabic Literature: An Introduction,* 2nd ed. (1974), a probing survey of 1,500 years of literature; SALIH J. ALTOMA, *Modern Arabic Literature* (1975), a bibliography of 850 general and scholarly works covering 1800–1970.

Muḥammad. *Biographical works:* W. MONTGOMERY WATT, *Muhammad at Mecca* (1953) and *Muhammad at Medina* (1956, reprinted 1977), a full-scale treatment, summarized in *Muhammad: Prophet and Statesman* (1961, reissued 1974), by the same author; FRANTS BUHL, . . . *Das Leben Muhammeds,* 3rd ed. (1961), a German translation of a work published in Danish in 1903 that is still considered reliable; TOR ANDRAE, *Mohammed: The Man and His Faith,* trans. by THEOPHIL MENZEL (1936, reprinted 1971), chiefly concerned with religious aspects; IBN HISHĀM, *The Life of Muhammad: A Translation of Isḥāq's Sīrat rasūl Allāh* (1955, reissued 1967), the primary Arabic biography; NABIA ABBOTT, *Aishah: The Beloved of Mohammed* (1942, reprinted 1973), a scholarly work; JOHN BAGOT GLUBB, *The Life and Times of Muhammad* (1970), a popular account based on the author's familiarity with Arab life.

Primary sources: The Qur'ān, of course, contains basic contemporary materials on Muḥammad, but it is difficult to assess them without broader historical knowledge. The vast collections of Traditions (Ḥadīth), or anecdotes about Muḥammad's words and deeds, are historically disputable and, besides, seldom tell us anything significant about Muḥammad's career. The main sources of historical value are the early biographies (8th–9th century), especially the *Sīrah* of IBN ISḤĀQ, as adapted by IBN HISHĀM, and the *Maghāzi* ("Expeditions") of AL-WAQIDI, together with the supplementary materials recorded by his associate IBN SA‘D. The last item contains much material about the Companions (persons in contact with Muḥammad) and thus about the Prophet's relation to and work with them. Some contemporary documents are preserved in the early biographical works, the most important being the so-called "Constitution of Medina." The latter is in Guillaume's translation of Ibn Isḥāq cited above; other documents are in Watt's *Muhammad at Medina.*

The Qur'ān. The basic work is T. NÖLDEKE, *Geschichte des Qorāns* (1860, 2nd ed. by FRIEDRICH SCHWALLY 1909–38, reprinted 1970). Less comprehensive but more modern are RICHARD BELL, *Introduction to the Qur'ān,* new ed., rev. and enl. by W. MONTGOMERY WATT (1970); and RÉGIS BLACHÈRE, *Introduction au Coran,* 2nd ed. (1977). The history of Qur'ānic interpretation is set forth in IGNÁC GOLDZIHER, *Die Richtungen der islamischen Koranauslegung* (1920, reprinted 1970). It should be supplemented by JOHANNES M.S. BALJON, *Modern Muslim Koran Interpretation, 1880–1960* (1961, reprinted 1968). ARTHUR JEFFERY, *The Qur'ān as Scripture* (1952, reprinted 1980), deals with the Qur'ān's view of its own function. KENNETH CRAGG, *The Event of the Qur'ān* (1971), deals with the meaning of the Qur'ān in Islāmic life. English translations include *The Meaning of the Glorious Qur'an: Text and Explanatory Translation,* by MUHAMMAD M. PICKTHALL (1938, reissued 1977); *The Koran Interpreted* by ARTHUR J. ARBERRY (1964), which is well known for its literary qualities; and HELMUT GÄTJE, *The Qur'ān and Its Exegesis: Selected Texts with Classical and Modern Muslim Interpretations,* trans. and ed. by ALFORD T. WELCH (1977).

Ḥadīth. J. ROBSON, "Hadīth," in *The Encyclopaedia of Islam,* new ed., vol. 3, pp. 23–28 (1971), and TH.W. JUYNBOLL, "Hadīth," in *The Encyclopaedia of Islam,* vol. 2, pp. 189–194 (1927), two important summaries with extensive bibliographies; ALFRED GUILLAUME, *The Traditions of Islam* (1924, reprinted 1980), a general introduction serviceable for a first study; IGNÁC GOLDZIHER, *Études sur la tradition islamique,* ed. by LÉON BERCHER (1952), a French trans. of the major part of vol. 2 of *Muhammedanische Studien,* 2 vol. (1888–90, Eng. trans. by C.R. BARBER and S.M. STERN, *Muslim Studies,* ed. by S.M. STERN, 1967–71), a classic work on the early development of Ḥadīth, reflecting the early history of religious ideas in Islām; MAULANA MUHAMMAD ALI, *A Manual of Hadith* (1944), a general selection, mainly from al-Bukhārī, in Arabic and English; A.J. WENSINCK, *A Handbook of Early Muhammadan Tradition* (1927, reprinted 1971), an alphabetical arrangement by a great Dutch scholar; MUHAMMAD Z. SIDDIQI, *Ḥādīth Literature: Its Origin, Development, Special Features and Criticism* (1961), an Asian Muslim's presentation.

Theology and philosophy. FRANZ ROSENTHAL (ed.), *The Classical Heritage in Islam,* trans. from the German by EMILE MARMORSTEIN and JENNY MARMORSTEIN (1975), and RICHARD WALZER, *Greek into Arabic: Essays on Islamic Philosophy* (1962, reissued 1970), deal with the Greek and Hellenistic background and its appropriation. PARVIZ MOREWEDGE (ed.), *Islamic Philosophical Theology* (1979), is a major contribution by nine internationally known authorities written for advanced students; W. MONTGOMERY WATT, *The Formative Period of Islamic Thought* (1973), is a study of the evolution of theological thought in the 300 years after Muḥammad's death, and his *Free Will and Predestination in Early Islam* (1948, reissued 1972), is an excellent treatment of the formative period of Islāmic theology; ASAF A.A. FYZEE (ed. and trans.), . . . *a Shī‘ite Creed* (1942), is an annotated translation of a standard Shī‘ite creed by Ibn Bābawayh; LOUIS GARDET and M.-M. ANAWATI, *Introduction à la théologie musulmane,* 2nd ed. (1970), is a comprehensive handbook on Sunnī theology; and A.J. WENSINCK, *The Muslim Creed* (1932, reprinted 1965), discusses the background and development of Sunnī doctrines. The theology of the Shī‘ah is given a prominent place in HENRY CORBIN, *Histoire de la philosophie islamique* (1964–); and its early development is discussed by WILFERD MADELUNG in both *Der Imam al-Qāsim ibn Ibrāhīm und die Glaubenslehre der Zaiditen* (1965), and "Imamism and Mu‘tazilite Theology," in *Le Shī‘isme imâmite,* pp. 13–30 (1970). M.M. SHARIF (ed.), *A History of Muslim Philosophy,* 2 vol. (1963–66), is a comprehensive collective work on the history of Islāmic philosophy and related subjects; it is especially useful for the later medieval and modern periods. MAJID FAKHRY, *A History of Islamic Philosophy* (1970), is a general history. FAZLUR RAHMAN discusses the development of the later synthesis between mysticism and philosophy in "Dream, Imagination and ‘Ālam al-Mithāl," *Islamic Studies,* vol. 3, no. 2, pp. 167–180 (June 1964) in the introduction to *Selected Letters of Shaikh Aḥmad Sirhindī* (1968), and in "The Eternity of the World and the Heavenly Bodies in Post-Avicennan Philosophy," in GEORGE F. HOURANI (ed.), *Essays on Islamic Philosophy and Science* (1975), a collection representing recent trends in interpreting Islāmic philosophy.

Islāmic mysticism. *Introductory works:* ARTHUR H. PALMER (comp.), *Oriental Mysticism: A Treatise on Sufiistic and Unitarian Theosophy of the Persians,* 2nd ed. by ARTHUR J. ARBERRY (1938, reprinted 1974), an exposition of later mystical ideas; ANNEMARIE SCHIMMEL, *Mystical Dimensions of Islam* (1975), a multifaceted, introductory study of Ṣūfism; REYNOLD A. NICHOLSON, *The Mystics of Islam* (1914, reprinted 1975), a very readable introduction to classical Ṣūfism and Ṣūfī poetry; ARTHUR J. ARBERRY, *Sufism: An Account of the Mystics of Islam* (1950), a historical survey of classical Ṣūfism; G.-C. ANAWATI and LOUIS GARDET, *Mystique musulmane,* 3rd ed. (1976), an excellent study of the major trends and leading personalities in classical Ṣūfism; ROBERT C. ZAEHNER, *Hindu and Muslim Mysticism* (1960, reissued 1969), a thought-provoking study of the possible relations between Indian and early Muslim mysticism.

History: MARGARET SMITH, *Rābi'a the Mystic & Her Fellow-Saints in Islam: Being the Life and Teachings of Rābi'a al-'Adawiyya al-Qaysiyya of Basra, Together with Some Account of the Place of the Women Saints in Islam* (1928, reprinted 1977), the first study of the herald of mystical love in Islām; JOSEPH VAN ESS, *Die Gedankenwelt des Ḥārit al-Muḥāsibī anhand von übersetzungen aus seinen Schriften dargestellt und erläutert* (1961), an excellent introduction to the theology and psychology of early mystical thought in Islām; LOUIS MASSIGNON, *La Passion de Husayn ibn Mansûr Hallâj: martyr mystique de l'Islam,* new ed. 4 vol. (1975), an indispensable source book for the history of Ṣūfism in the classical period; ANNEMARIE SCHIMMEL, *Al-Halladsch, Märtyrer der Gottesliebe* (1968), a German translation of parts of Ḥallāj's poetry and prose, and a study of his influence on the literatures of the different Islāmic peoples; SERGE DE LAUGIER DE BEAURECUEIL, *Khwādja 'Abdullāh Anṣārī (396–481 H./1006–1089): Mystique Hanbalite* (1965), a biography of the author of the beautiful Persian *munājāt* (prayers) and other mystical books; A.J. WENSINCK, *La Pensée de Ghazzālī* (1940), a short and reliable introduction to Ghazālī's thought; JOHN A. SUBHAN, *Sufism: Its Saints and Shrines* (1938, reissued 1978), a useful survey of the later development of Islāmic mysticism.

Ṣūfī literature: HELMUT RITTER, *Das Meer der Seele* (1955, reissued 1978), an exhaustive work on Farīd ud-Dīn 'Aṭṭār's thought as reflected in his mystical poetry; JALÁLU'DDIN RÚMÍ, *The Mathnawí,* ed. with critical notes, translation, and commentary by REYNOLD A. NICHOLSON, 8 vol. (1925–40), the encyclopaedia of mystical thought in the 13th century in masterly translation; H.T. SORLEY, *Shah Abdul Latīf of Bhit* (1940; reprinted 1966), a study of the greatest mystical poet of Sind.

Ṣūfī thought and practice: BENEDIKT REINERT, *Die Lehre vom Tawakkul in der klassischen Sufik* (1968), the first fundamental study of a single concept central to early Islāmic mysticism, built upon a critical analysis of all available sources; ARTHUR J. ARBERRY, *The Doctrine of the Ṣūfīs* (1935, reprinted 1977), a useful translation of Kalābādhi's *Kitāb at-ta'arruf,* one of the early treatises on Ṣūfī thought; *The Kashf al-Maḥjub: The Oldest Persian Treatise on Sufism* by ALI BIN UTHMAN al-HUJWIRI, trans. by REYNOLD A. NICHOLSON (1911, reprinted 1976), a masterly translation of the voluminous 11th-century account of Ṣūfī thought; G.-H. BOUSQUET (ed.), *Ih'yâ 'ouloûm ed-dîn; ou Vivification des sciences de la foi* (1955), an analytical index of the most widely read work on moderate mystical thought, prepared with the assistance of numerous scholars; CONSTANCE E. PADWICK, *Muslim Devotions* (1961), the only account of the popular mystically tinged piety of the Muslims as reflected in their prayer books; LALEH BAKHTIAR, *Sufi: Expressions of the Mystic Quest* (1976), discusses and shows through illustrations the Ṣūfī experience and its expression in the arts.

Theosophical Ṣūfism: A.E. AFFIFI, *The Mystical Philosophy of Muhyid Dín-Ibnul 'Arabí* (1939, reissued 1974), the first attempt, in a Western language, to systematize the pantheistic system of the 13th-century theosophist; HENRY CORBIN, *Creative Imagination in the Sūfism of Ibn 'Arabī,* trans. by RALPH MANHEIM (1970); REYNOLD A. NICHOLSON, *Studies in Islamic Mysticism* (1921, reissued 1978), a study of Abū Sa'īd and a discussion of Jīlī's Perfect Man and of Ibn al-Fāriḍ, with a superb translation of most of his odes.

Ṣūfī orders: OCTAVE DEPONT and XAVIER CAPPOLANI, *Les Confréries religieuses musulmanes* (1897), a comprehensive account of Ṣūfī brotherhoods; HANS J. KISSLING, "Die Wunder der Derwische," ZDMG (*Zeitschrift der deutschen morgenländischen Gesellschaft),* vol. 107, no. 2, pp. 348–361 (August 1957), a fully documented account of the kinds of miracles performed by dervishes; KHALIQ A. NIZAMI, *The Life and Times of Shaikh Faridùd-din Ganj-i-Shakar* (1955, reprinted 1973), a good survey of the life of one of the leading Chishtī saints in India; RENÉ BRUNEL, *Le Monachisme errant dans l'Islam: Sīdi Heddi et les Heddāwa* (1955), a penetrating study of a little known fraternity of dervishes in North Africa; JAMIL M. ABUN-NASR, *The Tijaniyya: A Sufi Order in the Modern World* (1965), a study of the development of political activities of this 19th-century order in the northern and western parts of Africa; J. SPENCER TRIMINGHAM, *The Sufi Orders in Islam* (1971), the first attempt to give a survey of all orders in Islām, and, as such, quite useful.

Islāmic law. A general survey of the Islāmic legal system, covering its historical development, jurisprudential theory, and the most important spheres of the substantive law, is contained in JOSEPH SCHACHT, *An Introduction to Islamic Law* (1964); JAMES N.D. ANDERSON, *Islamic Law in the Modern World* (1959, reprinted 1975); and NOEL J. COULSON, *History of Islamic Law* (1964, reprinted 1971). The reader is referred to the bibliographies of these books, particularly for the numerous articles written by James N.D. Anderson on developments in the law. JOSEPH SCHACHT, *Origins of Muhammadan Jurisprudence* (1950, reissued with corrections and additions, 1967), is a fundamental work of modern research on the early development of legal theory written by the pioneer scholar of this subject. A sound analysis of traditional legal theory is presented in ABDUR RAHIM, *Muhammadan Jurisprudence* (1911, reprinted 1981); and in *The Philosophy of Jurisprudence in Islam,* (1961), an Eng. trans. by FARMAT J. ZIADEH of the Arabic text of an outstanding Muslim jurist, Subhi Mahmassani. MAJID KHADDURI and HERBERT J. LIEBESNY (eds.), *Law in the Middle East* (1955), includes chapters by Muslim scholars and Western Orientalists on the various spheres of substantive Islāmic law, traditional and modern. ASAF A.A. FYZEE, *Outlines of Muhammadan Law,* 4th ed. (1974), is a standard text dealing with Islāmic law as it is applied in India and Pakistan. NORMAN ANDERSON, *Law Reform in the Muslim World* (1976), is a comparative study of the history, philosophy, and achievements of legal reform. *The Encyclopaedia of Islam* (1913–42; new ed., 1960–), contains numerous articles on individual legal topics.

Islāmic myth and legend. TOR ANDRAE, *Die Person Muhammeds in Lehre und Glauben seiner Gemeinde* (1917), on the development of Muḥammad-mysticism; ISRAEL FRIEDLÄNDER, *Die Chadhirlegende und der Alexander-Roman* (1913), on the relation between the Alexander romance and the figure of Khiḍr; MAX J.H. HORTEN, *Die religiöse Gedankenwelt der gebildeten Muslime in heutigen Islam* (1916), an account of popular Islām, and *Die religiöse Gedankenwelt des Volkes im heutigen Islam,* 2 pt. (1917–18), an account of the ideas of educated people in Islām; A.J. WENSINCK, "The Ocean in the Literature of the Western Semites," *Verhandelingen der Koninklijke Akademie van Wetenschappen,* vol. 19, no. 2 (1918), and "The Ideas of the Western Semites Concerning the Navel of the Earth," *ibid.,* vol. 17, no. 1 (1916); SEYYED H. NASR, *Three Muslim Sages* (1964, reissued 1976), an account of the theories of Suhrawardī al-Maqtūl and Ibn 'Arabī; JOSEPH HOROWITZ, "The Growth of the Mohammed Legend," *Moslem World,* vol. 10, no. 1, pp. 49–58 (January 1920), stresses the haggadic influences; WALTHER EICKMANN, *Angelologie und Dämonologie des Korans . . .* (1908), a study of the Qur'ānic concepts of angels and demons; ERNST A. ZBINDEN, *Die Djinn des Islam und der altorientalische Geisterglaube* (1953), a study of the different types of spirits in Islāmic folklore and tradition; RUDOLF KRISS and HUBERT KRISS-HEINRICH, *Volksglaube im Bereich des Islam,* 2 vol. (1960–62), useful studies in Islāmic folklore, with extensive bibliographies; TAUFIC CANAAN, *Mohammedan Saints and Sanctuaries in Palestine* (1927), on Palestinian folklore; articles in the *Shorter Encyclopaedia of Islam* (1953), an authoritative collection of information, each article furnished with an extensive bibliography.

Islāmic Arts

The vast populations of the Middle East and elsewhere that adopted the Islāmic faith from the 7th century onward have created such an immense variety of literatures, performing arts, visual arts, and music that it virtually defies any comprehensive definition. In the narrowest sense, the arts of the Islāmic peoples might be said to include only those arising directly from the practice of Islām; more commonly, however, the term is extended to include all of the arts produced by Muslim peoples, whether connected with their religion or not. In this article, the subject includes the arts created in pre-Islāmic times by Arabs and other peoples in Asia Minor and North Africa who eventually adopted the Islāmic faith. On the other hand, arts produced in cultural areas that were only partially Muslim, such as South Asia, Southeast Asia, and Central Asia, are discussed primarily in articles on arts of those regions.

This article is divided into the following sections:

General considerations

It is difficult to establish a common denominator for all of the artistic expressions of the Islāmic peoples. Such a common denominator would have to be meaningful for miniature painting and historiography, for a musical mode and the form of a poem. The relationship between the art of the Islāmic peoples and its religious basis is anything but direct.

The prohibition on representation

Like most prophetic religions, Islām is not conducive to fine arts. Representation of living beings is prohibited—not in the Qurʾān but in the prophetic tradition. Thus, the centre of the Islāmic artistic tradition lies in calligraphy, a distinguishing feature of this culture, in which the word as the medium of divine revelation plays such an important role. Representational art was found, however, in some early palaces and "at the doors of the bathhouses," according to later Persian poetry. After the 13th century a highly refined art of miniature developed, primarily in the non-Arab countries; it dwells, however, only rarely upon religious subjects. The typical expression of Muslim art is the arabesque, both in its geometric and in its vegetabilic form—one leaf, one flower growing out of the other, without beginning and end and capable of almost innumerable variations—only gradually detected by the eye—which never lose their charm. An aversion to empty spaces distinguishes that art; neither the tile-covered walls of a mosque nor the rich imagery of a poem allows an unembellished area; and the decoration of a carpet can be extended almost without limit.

The centre of Islāmic religion is the clean place for prayer, enlarged into the mosque, which comprises the community and all its needs. The essential structure is similar throughout the Muslim world. There are, of course, period and regional differences—large, wide court mosques of early times; court mosques, with big halls, of Iran and adjacent countries; central buildings with the wonderfully shaped domes of the Ottoman Empire. The implements, however, are the same: a niche (*miḥrāb*)—pointing to Mecca—made of wood, marble, mosaic, stone, tiles; a small pulpit for the Friday sermon; minarets, locally differently shaped but always rising like the call to prayer that is uttered from their tops; the wooden carved stands for the Qur'ān, which is to be written in the most perfect form; sometimes highly artistic lamps (made in Syria and proverbially mentioned all over the Muslim world); perhaps bronze candlesticks, with inlaid ornaments; and rich variations of the prayer mats. If any decoration was needed, it was the words of God, beautifully written or carved in the walls or around the domes. At first connected with the mosques and later independent of them are schools, mausoleums, rooms for the students, and cells for the religious masters.

The poetry of the Arabs consisted in the beginning of praise and satirical poems thought to be full of magic qualities. The strict rules of the outward form of the poems (monorhyme, complicated metre) even in pre-Islāmic times led to a certain formalism and encouraged imitation.

Goethe's statement that the stories of *The Thousand and One Nights* have no goal in themselves shows his understanding of the character of Arabic belles lettres, contrasting them with the Islāmic religion, which aims at "collecting and uniting people in order to achieve one high goal." Poets, on the other hand, rove around without any ethical purpose, according to the Qur'ān. For many pious Muslims, poetry was something suspect, opposed to the divine law, especially since it sang mostly of forbidden wine and of free love. The combination of music and poetry, as practiced in court circles and among the mystics, has always aroused the wrath of the lawyer divines who wielded so much authority in Islāmic communities. This opposition may partly explain why Islāmic poetry and fine arts took refuge in a kind of unreal world, using fixed images that could be correctly interpreted only by those who were knowledgeable in the art.

The ambiguity of Persian poetry, which oscillates between the worldly, the divine, and often the political level, is typical of Islāmic writings. Especially in Iran and the countries under its cultural influence, this kind of poetry formed the most important part of literature. Epic poetry of all kinds developed exclusively outside the Arabic-speaking countries; Western readers look in vain for an epical structure in such long poems (as in the case of the prose-romances of the Arabs) and find, instead, a rather aimless representation of facts and fictions. A similar characteristic even conditions innumerable historical works in Arabic, Persian, and Turkish, which, especially in classical times, contain much valuable information, put together without being shaped into a real work of art; only rarely does the historian or philosopher reach a comprehensive view. The first attempt at a philosophy of history, Ibn Khaldūn's *Muqaddimah,* in the 14th century, was rarely studied by his Arab compatriots.

The accumulation of large amounts of material, which is carefully organized up to the present, seems typical of all branches of Islāmic scholarship, from theology to natural sciences. There are many minute observations and descriptions but rarely a full view of the whole process. Later, especially in the Persian, Turkish, and Indo-Muslim areas, a tendency to overstress the decorative elements of prose is evident; and the contents even of official chronicles are hidden behind a network of rhymed prose, which is difficult to disentangle.

The characteristic lack of structure

This tendency is illustrated in all branches of Islāmic art: the lack of "architectural" formation. Instead, there is a kind of carpet-like pattern; the Arabic and Persian poem is, in general, judged not as a closed unity but rather according to the perfection of its individual verses. Its main object is not to convey a deep personal feeling but to perfect to the utmost the traditional rules and inherited metaphors, to which a new image may sometimes be added; thus the personality of the poet becomes visible only through the minimal changes of expression and rhythm and the application of certain preferred metaphors, just as the personality of the miniature painter can be detected by a careful observation of details, of his way of colouring a rock or deepening the shade of a turban. The same holds true for the arabesques, which were developed according to a strict ritual to a mathematical pattern and were refined until they reached a perfection of geometrical complicated figures, as in the dome of the Karatay Medrese in Konya (1251); it corresponds both to the most intricate lacelike Kūfic inscriptions around this dome and to the poetical style of Jalāl ad-Dīn ar-Rūmī, who wrote in that very place and during those years. His immortal mystical poems comprise thousands of variations on the central theme of love. Although such a perfect congruency of poetry and fine arts is not frequently found, the precept about Persian art that "its wings are too heavy with beauty" can also be applied to Persian poetry. Thus, the tile work of a Persian mosque, which combines different levels of arabesque work with different styles of writing, is reminiscent of the way Persian poetry combines at least two levels of reality. And a perfect harmony is reached in some of the miniature manuscripts of Iran, Muslim India, or Ottoman Turkey, which, in their lucid colours and fine details of execution, recall both the perfection of the calligraphy that surrounds them on delicate paper and the subtlety of the stories or poems that they accompany or illustrate.

Those accustomed to the Western ideals of plasticity or form in the fine arts and literature or to the polyphonic interweaving of melodic lines in music have some difficulties in appreciating this art. The palaces seem to be without a fixed architectural plan; rooms and gardens are simply laid out according to daily needs. The historian offers an astounding amount of detailed reports and facts but with no unifying concept. The Muslim writer prefers this carpet-like form; he adds colour to colour, motif to motif, so that the reader only understands the meaning and end of the whole web from a certain distance. Music, differentiated as it may be in the countries between Morocco and India, follows the same model: variations of highest subtlety on a comparatively simple given subject or theme.

Man, a puppet; God, the only actor

Drama and opera in the Western sense did not develop in the Islāmic countries until the 19th century; and the art of the novel is a very recent development. There was no reason for drama: in the Muslim perception God is the only actor who can do whatever he pleases, whose will is inscrutable. Man is, at best, a puppet on a string, behind whose movement those with insight detect the hand of the play master; neither is the problem of personal guilt and absolution posed as it is in the West, nor is a catharsis, or purging of emotion, needed through drama. The atomist theory, widely accepted in Islām since the 10th century, leaves no room for a "dramatic" movement; it teaches that God creates everything anew in every moment, and what is called a "law of nature" is nothing but God's custom, which he can interrupt whenever he pleases.

It is true that certain other forms are found in the more folkloristic arts of Islām. Every region has produced poetry, in regional languages, that is more lively and more realistic than the classical court poetry; but such poetry tends to become restricted to certain fixed forms that can be easily imitated. Attempts at drama in Islām come from these more popular spheres in Iran (and, rarely, in Lebanon and Iraq), where the tragic events of the murder of Ḥusayn (680) at Karbalā' were dramatized in strange forms, using the vocabulary of traditional Persian poetry and theology. Thus, strangely hybrid forms emerge in the Islāmic arts, highly interesting for the historian of religion and the student of literature but not typical of the classical Islāmic ideals. Popular illustrations of tales and legends and those of some of the Shī'ah heroes are similarly interesting but atypical. In modern times, of course, there have been imitations of all forms of Western literary and visual arts: paintings in the Impressionist or Cubist style, the use of free verse instead of the stern classical forms; and novels, dramas, motion pictures, and music combining Western and Eastern modes. Belief in the Qur'ānic dictum "Whatever is on earth will perish save His face" discouraged artistic endeavour on a large scale; but the Prophetic tradition "Verily God is beautiful and loves beauty" has inspired numberless artists and artisans, writers and poets, musicians, and mystics to develop their arts and crafts as a reflection of that divine beauty. A theory of aesthetics comprising the various artistic expressions of the Muslim peoples has yet to be written. Although there have been a number of studies in literary criticism, the formal indebtedness of some of the best modern poets and painters to the Islāmic heritage has never been studied in full.

It is notable that the arts of the Islāmic peoples have had relatively little impact on other cultures, certainly far less than their artistic merit would appear to warrant.

Islāmic art in the West

Europe has known art objects of Islāmic origin since the early Middle Ages, when they were brought home by the crusaders or manufactured by the Arabs in Sicily and Spain. Much admired and even imitated, they formed part of the material culture in those times, so much so that even the coronation robes of the German emperor were decorated with an Arabic inscription. At the same time, Islāmic motives wandered into the belles lettres of Europe, and Islāmic scientific books formed a basis for the development of Western science. Islāmic culture as such, however, was rather an object of hatred than of admiration; a more objective appreciation of both the works of art and of literature did not start until the mid-17th century, when travelers told of the magnificent buildings in Iran and Mughal India, and the first works from Persian literature were translated, influencing German classical literature. Indian miniatures inspired Rembrandt, just as European paintings were imitated by Islāmic, especially Mughal, artists. Persian carpets were among the most coveted gifts for princes and princesses.

A bias against the cultures of the East persisted, however, until after the 18th-century Age of Enlightenment; the indefatigable work of the British scholars at Fort William at Calcutta brought new literary treasures to Europe, where they were studied carefully by specialists in the emerging field of Islāmic studies. Poets such as Goethe in Germany in the early 19th century paved the way for a deeper understanding of Islāmic poetry. Islāmic literatures, however, continue to be known to the larger Western public almost exclusively by *The Thousand and One Nights,* or *The Arabian Nights' Entertainment* (translated first in the early 18th century), Omar Khayyam's *robāʿīyāt,* and the lyrics of Ḥāfeẓ. Even experts who are aware of the immense wealth of the literatures in the different Islāmic languages (such as Arabic, Persian, Turkish, and Urdu) until now have rarely appreciated the literatures from an aesthetic viewpoint; rather, they have used them as a source for lexicography and for philological and historical research. The situation in Islāmic fine arts and architecture is similar. Although the beauty of the Alhambra, for example, had already inspired European scholars and artists in the early 19th century, a thorough study of Islāmic art as an independent field began only in the 20th century. There was even less interest in the music of Islāmic peoples, the arabesque-like uniformity of which seems strange to Western ideals of harmony.

Islāmic literatures

NATURE AND SCOPE

It would be almost impossible to make an exhaustive survey of Islāmic literatures. There are so many works, of which hundreds of thousands are available only in manuscript, that even a very large team of scholars could scarcely master a single branch of the subject. Islāmic literatures, moreover, exist over a vast geographical and linguistic area, for they were produced wherever the Muslims went, pushing out from their heartland in Arabia through the countries of the Near and Middle East as far as Spain, North Africa, and, eventually, West Africa. Iran (Persia) is a major centre of Islām, along with the neighbouring areas that came under Persian influence, including Turkey and the Turkic-speaking peoples of Central Asia. Many Indian vernaculars contain almost exclusively Islāmic literary subjects; there is an Islāmic content in the literature of Malaysia and in that of some East African languages, including Swahili. In many cases, however, the Islāmic content proper is restricted to religious works—mystical treatises, books on Islāmic law and its implementation, historical works praising the heroic deeds and miraculous adventures of earlier Muslim rulers and saints, or devotional works in honour of the prophet Muḥammad.

The vast majority of Arabic writings are scholarly—the same, indeed, is true of the other languages under discussion. There are superb, historically important translations made by medieval scholars from Greek into Arabic; historical works, both general and particular; a range of religiously inspired works; books on grammar and on stylistics, on ethics and on philosophy. All have helped to shape the spirit of Islāmic literature in general, and it is often difficult to draw a line between such works of "scholarship" and works of "literature" in the narrower sense of that term. Even a strictly theological commentary can bring about a deeper understanding of some problem of aesthetics. A work of history composed in florid and "artistic" language would certainly be regarded by its author as a work of art as well as of scholarship, whereas the grammarian would be equally sure that his keen insights into the structure of Arabic grammar were of the utmost importance in preserving that literary beauty in which Arabs and non-Arabs alike took pride.

Pride in literary beauty

In this treatment of Islāmic literatures, however, the definition of "literature" is restricted to poetry and belles lettres, whether popular or courtly in inspiration. Other categories of writing will be dealt with briefly if these shed light on some peculiar problem of literature.

The range of Islāmic literatures. Although Islāmic literatures appear in such a wide range of languages and in so many different cultural environments, their unity

is safeguarded by the identity of the basic existential experience, by the identity of the fundamental intellectual interests, by the authoritativeness of certain principles of form and presentation, not to mention the kindred political and social organization within which those peoples aspire to live.

Arabic: language of the Qurʾān. The area of Islāmic culture extends from western Africa to Malaysia, Indonesia, and the Philippines; but its heartland is Arabia, and the prime importance and special authority of the Arabic language was to remain largely unquestioned after the spread of Islām. The Arabic poetry of pre-Islāmic Arabia was regarded for centuries afterward as the standard model for all Islāmic poetic achievement, and it directly influenced literary forms in many non-Arab literatures. The Qurʾān, Islām's sacred scripture, was accepted by pious Muslims as God's uncreated word and was considered to be the highest manifestation of literary beauty. A whole literature defended its inimitability (*iʿjāz*) and unsurpassable beauty. Because it was God's own word, the Qurʾān could not legitimately be translated into any other language; the study of at least some Arabic was therefore required of every Muslim. Arabic script was used by all those peoples who followed Islām, however much their own languages might differ in structure from Arabic. The Qurʾān became the textbook of the Muslims' entire philosophy of life; theology, lexicography, geography, historiography, and mysticism all grew out of a deep study of its form and content; and even in the most secular works there can be found allusions to the holy book. Its imagery not unexpectedly permeates all Islāmic poetry and prose.

Allusions to the Qurʾān in all literary genres

Between the coming of Islām in the 7th century and the 11th, a great deal of poetry and prose in Arabic was produced. One branch of literature in Spain and North Africa matured in perfect harmony with the classical ideals of the Muslim East although its masters, during the 11th and 12th centuries, invented a few strophic forms unknown to classical Arabic poetry. In modern times, North African Muslim literature—mainly from Algeria and Morocco—often uses French as a means of expression, since the tradition of Arabic writing was interrupted by the French occupation in the 19th century and has had to be built up afresh.

Persian. In 641 the Muslims entered Iran, and Persian influence on literary taste becomes apparent in Arabic literature from the mid-8th century onward. Many stories and tales were transmitted from, or through, Iran to the Arab world and often from there to western Europe. Soon Iran could boast a large literature in its own tongue. Persian literature was more varied in its forms and content than that written in classical Arabic. Although Persian adopted many of the formal rules of the Arabic language (including prosody and rhyme patterns), new genres, including epic poetry, were introduced from Iran. The lyric, elegant and supple, also reached its finest expression in the Persian language.

South Asian. Persian culture was by no means restricted to Iran itself. Northwestern India and what is now Pakistan became a centre of Islāmic literature as early as the 11th century, with Delhi and Agra being of special importance. It was to remain a stronghold of Muslim cultural life, which soon also extended to the east (Bengal) and south (Deccan). Persian remained the official language of Muslim India until 1835, and not only its poetry but even its historiography was written in the high-flown manner that exemplified the Persian concept of fine style. Muslim India can further boast a fine heritage of Arabic poetry and prose (theological, philosophical, and mystical works).

At various times in its history the Indian subcontinent was ruled by princes of Turkish origin (indeed, the words *Turk* and *Muslim* became synonymous in some Indian languages). The princes surrounded themselves with a military aristocracy of mainly Turkish extraction, and thus a few poetical and prose works in Turkish were written at some Indian courts. In various regions of the subcontinent an extremely pleasing folk literature has flourished throughout the ages: Sindhi in the lower Indus Valley, for example, and Punjabi in the Punjab are languages rich in an emotional poetry that uses popular metres and forms. At the Indo-Iranian border the oldest fragments of the powerful Pashto poetry date from the Middle Ages. The neighbouring Baluchi poetry consists largely of ballads and religious folksongs. All the peoples in this area have interpreted Islāmic mysticism in their own simple, touching imagery. In the east of the subcontinent, Bengali Muslims possess a large Islāmic literary heritage, including religious epics from the 14th and 15th centuries and some lovely religious folksongs. The achievements of modern novelists and lyric poets from Bangladesh are impressive. To the north, where Islām came in the 14th century, a number of classical themes in Islāmic lore were elaborated in Kashmiri lyric and epic poetry. To the south, an occasional piece of Islāmic religious poetry can be found even in Tamil and Malayalam. Some fine Muslim short stories have been produced in modern Malayalam.

Islāmic mystical thought in folk literature

Urdu, now the chief literary language of Muslim India and Pakistan, borrowed heavily from Persian literature during its classical period in the 18th century. In many writings only the verbs are in Urdu, the rest consisting of Persian constructions and vocabulary; and the themes of traditional Urdu literature were often adapted from Persian. Modern Urdu prose, however, has freed itself almost completely from the past, whereas in poetry promising steps have been taken toward modernization of both forms and content (see SOUTH ASIAN ARTS).

Turkish. An elaborate "classical" style developed in Turkish after the 14th century, reaching its peak in the 17th. Like classical Urdu, it was heavily influenced by Persian in metrics and vocabulary. Many exponents of this "high" style came from the Balkan provinces of the Ottoman Empire. On the other hand, a rich and moving folk poetry in popular syllable-counting metres has always flourished among the Turkish population of Anatolia and Rumelia. The mystical songs of their poet Yunus Emre (died *c.* 1321) contributed greatly toward shaping this body of literature, which was preserved in the religious centres of the Ṣūfī orders of Islām. From this folk tradition, as well as from Western literature, modern Turkish literature has derived a great deal of its inspiration.

Turkic languages. A great deal of the Muslim literature of Central Asia is written in Turkic languages, which include Uzbek, Tatar, and Kyrgyz. Its main cultural centres (Samarkand, Bukhara, Fergana) became part of the Muslim empire after 711. Central Asia was an important centre of Islāmic learning until the Tsarist invasions in the 1870s, and the peoples of this region have produced a classical literature in Arabic. Many of the most famous Arabic and Persian scholars and poets writing in the heyday of Muslim influence were Central Asians by birth. Central Asians also possess a considerable literature of their own, consisting in large part of epics, folktales, and mystical "words of wisdom." The rules of prosody which hold for Arabic and Persian languages have been deliberately imposed on the Turkic languages on several occasions, notably by ʿAlī Shīr Navāʾī (died 1501), a master of Chagatai poetry and prose in Herāt, and by Bābur (died 1530), the first Mughal emperor in India. Tajik literature is basically Persian, both as it is written today in Tajikistan and as it existed in earlier forms, when it was indistinguishable from classical Persian. After the Russification of the country, and especially after the 1917 Revolution, a new literature emerged that was part and parcel of the former Soviet Union's literature. The same can be said, by and large, about the literatures of other Muslim Turkic peoples of Central Asia.

Other languages. Smaller fragments of Islāmic literature, in Chinese, are found in China (which has quite a large Muslim population) and in the Philippines. The literary traditions of Indonesia and of Malaysia, where the religion of Islām arrived long ago, are also worth noting. Historical and semimythical tales about Islāmic heroes are a feature of the literature in these areas, a fact of immense interest to folklorists.

Contact with Islām and its "written" culture also helped to preserve national idioms in many regions. Often such idioms were enriched by Arabic vocabulary and Islāmic concepts. The leaders of the Muslims in such areas in northern Nigeria, for example, preferred to write poetry and chronicles in Arabic, while using their mother tongue for more popular forms of literature (see AFRICAN ARTS). Of particular interest in this connection is Kurdish literature, which has preserved in an Iranian language several important, popular heterodox texts and epics.

Islāmic literatures and the West. Small fragments of Arabic literature have long been known in the West. There were cultural interrelations between Muslim Spain (which, like the Indus Valley, became part of the Muslim empire after 711) and its Christian neighbours, and this meant that many philosophical and scientific works filtered through to western Europe. It is also likely that the poetry of Muslim Spain influenced the growth of certain forms of Spanish and French troubadour poetry and provided an element, however distorted, for medieval Western romances and heroic tales.

First Western studies

Investigation of Oriental literatures by Western scholars did not begin until the 16th century in the Netherlands and England. First attempts toward an aesthetic understanding of Arabic and Persian poetry came even later: they were made by the British Orientalists of Fort William, Calcutta, and by German pre-Romantics of the late 18th century. In the first half of the 19th century the publication of numerous translations of Oriental poetry, especially into German, began to interest some Europeans. The poetical translations from Arabic, Persian, and Sanskrit made by the German Orientalist and poet Friedrich Rückert can scarcely be surpassed, either in accuracy or in poetical mastery. The Persian poet Ḥāfeẓ became well known in German-speaking countries, thanks to Johann Wolfgang von Goethe's enchanting poems, *West-östlicher Divan* (1819), a collection which was the first response to Persian poetry and the first aesthetic appreciation of the character of Oriental poetry by an acknowledged giant of European literature. An "Orientalizing style," which employed Arabo-Persian literary forms such as the *ghazal* (a short, graceful poem with monorhyme), became fashionable at times in Germany. Later, Edward FitzGerald aroused new interest in Persian poetry with his free adaptations of Omar Khayyam's *robāʿīyāt* (*The Rubáiyát of Omar Khayyám*, 1859). The fairy tales known as *The Thousand and One Nights*, first translated in 1704, provided abundant raw material for many a Western writer's play, novel, story, or poem about the Islāmic East.

EXTERNAL CHARACTERISTICS

In order to understand and enjoy Oriental literature, the external characteristics of it have to be studied most carefully. The literatures of the Islāmic peoples are "intellectual"; in neither poetry nor prose are there many examples of subjective lyricism, as it is understood in the West. The principal genres, forms, and rules were inherited from pre-Islāmic Arabic poetry but were substantially elaborated afterward, especially by the Persians.

Rhyme and metre. Arabic poetry is built upon the principle of monorhyme, and the single rhyme, usually consisting in one letter, is employed throughout every poem, long or short. The structure of Arabic permits such monorhymes to be achieved with comparative ease. The Persians and their imitators often extended the rhyming part over two or more syllables (*radīf*) or groups of words, which are repeated after the dominant rhyming consonant. The metres are quantitative, counting long and short syllables (*ʿarūḍ*). Classical Arabic has 16 basic metres in five groupings; they can undergo certain variations, but the poet is not allowed to change the metre in the course of his poem. Syllable-counting metres, as well as strophic forms, are used in popular, or "low," poetry; only in post-classical Arabic were some strophic forms introduced into "high" poetry. Many modern Islāmic poets, from Pakistan to Turkey and North Africa, have discarded the classical system of prosody altogether. In part they have substituted verse forms imitating Western models such as strophic poems with or without rhyme; since about 1950 free verse has almost become the rule, although a certain tendency toward rhyming or to the use of alliterative quasi-rhymes can be observed.

Genres. The chief poetic genres, as they emerged according to traditional rules, are the *qaṣīdah*, the *ghazal*, and the *qiṭʿah;* in Iran and its adjacent countries there are, further, the *robāʿī* and the *mas̄navī*.

Qaṣīdah. The *qaṣīdah* (literally "purpose poem"), a genre whose form was invented by pre-Islāmic Arabs, has from 20 to more than 100 verses and usually contains an account of the poet's journey. In the classic pattern, the parts followed a fixed sequence, beginning with a love-poem prologue (*nasīb*), followed by a description of the journey itself, and finally reaching its real goal by flattering the poet's patron, sharply attacking some adversaries of his tribe, or else indulging in measureless self-praise. Everywhere in the Muslim world the *qaṣīdah* became the characteristic form for panegyric. It could serve for religious purposes as well: solemn praise of God, eulogies of the Prophet, and songs of praise and lament for the martyr heroes of Shīʿah Islām were all expressed in this form. Later, the introductory part of the *qaṣīdah* often was taken up by a description of nature or given over to some words of wisdom; or the poet took the opportunity to demonstrate his skill in handling extravagant language and to show off his learning. Such exhibitions were made all the more difficult because, though it varied according to the rank of the person to whom it was addressed, the vocabulary of each type of *qaṣīdah* was controlled by rigid conventions. This type of poetry, however, could obviously lend itself easily to empty verbosity or to pedantry.

Ghazal. The *ghazal* possibly originated as an independent elaboration of the *qaṣīdah*'s introductory section, and it usually embodies a love poem. Ideally, its length varies between five and 12 verses. It can be used either for religious or secular expression, the two often being blended indistinguishably. Its diction is light and graceful, its effect comparable to that of chamber music, whereas the *qaṣīdah*-writer employs, so to speak, the full orchestral resources.

The love poem

Qiṭʿah. Monorhyme is used in both the *qaṣīdah* and *ghazal.* But while these two forms begin with two rhyming hemistiches (half-lines of a verse), in the *qiṭʿah* ("section") the first hemistich does not rhyme, and the effect is as though the poem had been "cut out" of a longer one (hence its name). The *qiṭʿah* is a less serious literary form that was used to deal with aspects of everyday life; it served mainly for occasional poems, satire, jokes, word games, and chronograms.

Robāʿī. The form of the *robāʿī*, which is a quatrain in fixed metre with a rhyme scheme of *a a b a*, seems to go back to pre-Islāmic Persian poetical tradition. It has supplied the Persian poets with a flexible vehicle for ingenious aphorisms and similarly concise expressions of thought for religious, erotic, or skeptical purposes. The peoples who came under Persian cultural influence happily adopted this form.

Mas̄navī. Epic poetry was unknown to the Arabs, who were averse to fiction, whether it was expressed in poetry or in prose. The development of epic poetry was thus hindered, just as was the creation of novels or short stories. Nevertheless, *mas̄navī*—which means literally "the doubled one," or rhyming couplet, and by extension a poem consisting of a series of such couplets—became a favourite poetical form of the Persians and those cultures they influenced. The *mas̄navī* enabled the poet to develop the thread of a tale through thousands of verses. Yet even in such poetry, only a restricted number of metres was employed, and no metre allowed more than 11 syllables in a hemistich. Metre and diction were prescribed in accordance with the topic; a didactic *mas̄navī* required a style and metre different from a heroic or romantic one. The *mas̄navī* usually begins with a praise of God, and this strikes the keynote of the poem.

Other poetic forms. There is a variety of other forms that are more or less restricted to folk poetry, such as the *sīḥarfī* ("golden alphabet"), in which each line or each stanza begins with succeeding letters of the Arabic alphabet. In Muslim India the *bārāmāsa* ("12 months") is a sort of lovers' calendar in which the poet, assuming the role of a young woman of longing, expresses the lover's feelings in accord with the seasons of the year. Apart from these, later writers tried to develop strophic forms. Sometimes *ghazal*s with the same metre were bound together as "stanzas" to form a longer unit through the use of a linking verse. When the linking verse was recurrent, the poem was called a *tarjīʿ-band* (literally "return-tie"); when the linking verse was varied, the poem was called a *tarkīb-band* (literally "composite-tie"). True stanzas of varying lengths were also invented. Among these, mainly in Urdu and Turkish, a six-line stanza known as *musaddas* became the form used for the *marsīyeh* (dirge for the martyrs of Karbalāʾ). Because it had come to be associated with lofty feeling and serious thought, *musaddas* later was used for the first reformist modern poems.

The Arabs inherited a love for rhymed prose from pre-Islāmic Arabia. Although the extent of prose literature, even in the field of belles lettres, is very large, the novel and novella were introduced only after contact with European literatures.

Rhyming prose

Maqāmah. The most typical expression of the Arabic—and Islāmic—spirit in prose is the *maqāmah* (gathering, assembly), which tells basically simple stories in an extremely and marvelously complicated style (abounding in word plays, logographs, double entendre, and the like) and which comes closest to the Western concept of the short story.

The versatility and erudition of the classical *maqāmah* authors is dazzling, but the fables and parables that, during the first centuries of Islām, had been told in a comparatively easy flowing style, later became subject to a growing trend toward artificiality, as did almost every other literary genre, including expository prose. Persian historiographers and Turkish biographers, Indo-Muslim writers on mysticism and even on science all indulged in a style in which rhyme and rhetoric often completely obscured the meaning. It is only since the late 19th century that a matter-of-fact style has slowly become acceptable in literary circles; the influence of translations from European languages, the role of journalism, and the growing pride in a pure language freed from the cobwebs of the past worked together to make Islāmic languages more pliable and less artificial.

Imagery. In all forms of poetry and in most types of prose, writers shared a common fund of imagery that was gradually refined and enlarged in the course of time. The main source of imagery was the Qurʾān, its figures and utterances often divested of their sacred significance. Thus, the beautiful Joseph (*sūrah* 12) is a fitting symbol for the handsome beloved; the nightingale may sing the psalms of David (*sūrah* 21:79 *a.o*); the rose sits on Solomon's wind-borne throne (*sūrah* 21:81 *a.o*), and its opening petals can be compared to Joseph's shirt rent by Potiphar's wife (*sūrah* 12:25 ff.), its scent to that of Joseph's shirt, which cured blind Jacob (*sūrah* 12:94). The tulip reminds the poet of the burning bush before which Moses stood (*sūrah* 20:9 ff.), and the coy beloved refuses the lover's demands by answering, like God to Moses, "Thou shalt not see me" (*sūrah* 7:143); but her (or his) kiss gives the dying lover new life, like the breath of Jesus (*sūrah* 3:49). Classical Persian poetry often mentions knights and kings from Iran's history alongside those from Arabic heroic tales. The cup of wine offered by the "old man of the Magians" is comparable to the miraculous cup owned by the Iranian mythical king Jamshīd or to Alexander's mirror, which showed the marvels of the world; the nightingale may sing "Zoroastrian tunes" when it contemplates the "fire temple of the rose." Central scenes from the great Persian *maṣ̌navīs* contributed to the imagery of later writers in Persian-, Turkish-, and Urdu-speaking areas. Social and political conditions are reflected in a favourite literary equation between the "beautiful and cruel beloved" and "the Turk": since in Iran and India the military caste was usually of Turkish origin, and since the Turk was always considered "white" and handsome, in literary imagery he stood as the "ruler of hearts." Minute arabesque-like descriptions of nature, particularly of garden scenes, are frequent: the rose and the nightingale have almost become substitutes for mythological figures. The versatile writer was expected to introduce elegant allusions to classical Arabic and Persian literature and to folklore and to know enough about astrology, alchemy, and medicine to use the relevant technical terms accurately. Images inspired by the pastimes of the grandees—chess, polo, hunting, and the like—were as necessary for a good poem as were those referring to music, painting, and calligraphy. Similarly, allusions in poetic imagery to the Arabic letters—often thought to be endowed with mystical significance or magical properties—were very common in all Islāmic literatures. The poet had to follow strict rules laid down by the masters of rhetoric, rigidly observing the harmonious selection of similes thought proper to any one given sphere (four allusions to Qur'ānic figures, for example; or three garden images all given in a single verse). The poet was expected to invent new fantastic etiologies (*ḥosn-e taʿlīl*): he had to describe natural phenomena in some elegant and surprising metaphor. Thus, "The narcissus has strewn silver in the way of the bride rose . . ." means simply "The narcissus has withered"—for when the rose (dressed in red, like an Oriental bride) appears in late spring it is time for the narcissus to shed its white petals, just as people would shed silver coins in the way of a bridal procession.

Qualities expected of the versatile writer

Skills required of the writer. The writer was also expected to use puns and to play with words of two or more meanings. He might write verses that could provide an intelligible meaning even when read backward. He had to be able to handle chronograms, codes based on the numerical values of a phrase or verse, which, when understood, gave the date of some relevant event. Later writers sometimes supplied the date of a book's compilation by hiding a chronogram in its title. A favourite device in poetry was the "question and answer" form, employed in the whole poem, or only in chosen sections.

One was expected to show his talent at both improvisation and elaboration on any theme if he wished to attract the interest of a generous patron. His poetry was judged according to the perfection of its individual verses. Only in rare cases was the poem appreciated as a whole: the lack of coherent argument, which often puzzles the Western reader in *ghazal* poetry, is in fact deliberate.

It would be idle to look for the sincere expression of personal emotion in Arabic, Turkish, or Persian poetry. The conventions are so rigid that the reader is allowed only a rare glimpse into the poet's feelings. Indeed, such feelings were put through the sieve of intellect, and personal experiences were thereby transformed into arabesque-like work of artistry, if not art. In the hands of mediocre versifiers and prose writers, however, literature became mannered and completely artificial. The reader soon tires of the constantly recurring moon faces, hyacinth curls, ruby lips, and cypress statures (that is, tall and slender). Yet the great masters of poetry and rhetoric (who all have their favourite imagery, rhymes, and rhythmical patterns) will sometimes allow the patient reader a glimpse into their hearts by a slight rhythmical change or by a new way of expressing a conventional thought.

These are, of course, quite crude generalizations. Folk poetry, for instance, has to be judged by different standards, though even here conventional forms and inherited imagery make it, on the whole, more standardized than might be wished. Only in the 20th century has a complete break with classical ideals been made—sincerity instead of monotonous imitation, political and social commitment instead of empty panegyric, realism instead of escapism: these are the characteristic features of modern literatures of the Muslim countries.

HISTORICAL DEVELOPMENTS: PRE-ISLĀMIC LITERATURE

The first known poetic compositions of the Arabs are of such perfect beauty and, at the same time, are so conventionalized, that they raise the question as to how far back an actual poetic tradition does stretch. A great number of pre-Islāmic poems, dating from the mid-6th century, were preserved by oral tradition. The seven most famous pieces are *al-Muʿallaqāt* ("The Suspended Ones," known as *The Seven Odes*), and these are discussed more fully below. The term *muʿallaqāt* is not fully understood: later legend asserts that the seven poems had been hung in the most important Arab religious sanctuary, the Kaʿbah in Mecca, because of their eloquence and beauty and had brought victory to their authors in the poetical contests traditionally held during the season of pilgrimage. Apart from these seven, quite a number of shorter poems were preserved by later scholars. An independent genre in pre-Islāmic poetry was the elegy, often composed by a woman, usually a deceased hero's sister. Some of these poems, especially those by the poetess al-Khansā' (died after 630) are notable for their compact expressiveness.

The first poetry

Poetry. The poet (called a *shāʿir*, a wizard endowed with magic powers) was thought to be inspired by a spirit (*jinn, shayṭān*). The poet defended the honour of his tribe and perpetuated their deeds. Religious expression was rare in pre-Islāmic poetry. In the main it reflects the sense of fatalism that was probably needed if the harsh circumstances of Bedouin life in the desert were to be endured.

The most striking feature of pre-Islāmic poetry is the uniformity and refinement of its language. Although the various tribes, constantly feuding with one another, all spoke their own dialects, they shared a common language for poetry whether they were Bedouins or inhabitants of the small capitals of al-Ḥīrah and Ghassān (where the influence of Aramaic culture was also in evidence).

Arabic was even then a virile and expressive language, with dozens of synonyms for the horse, the camel, the lion, and so forth; and it possessed a rich stock of descriptive adjectives. Because of these features, it is difficult for foreigners and modern Arabs alike to appreciate fully the artistic qualities of early Arabic poetry. Imagery is precise, and descriptions of natural phenomena are detailed. The sense of universal applicability is lacking, however, and the comparatively simple literary techniques of simile and metaphor predominate. The imaginative power that was later to be the hallmark of Arabic poetry under Persian influence had not yet become evident.

The virility of Arabic

The strikingly rich vocabulary of classical Arabic, as well as its sophisticated structure, is matched by highly elaborate metrical schemes, based on quantity. The rhythmical structures were analyzed by the grammarian Khalīl of Basra (died *c.* 791), who distinguished 16 metres. Each was capable of variation by shortening the foot or part of it; but the basic structure was rigidly preserved. One and the same rhyme letter had to be maintained throughout the poem. (The rules of rhyming are detailed and very complicated but were followed quite strictly from the 6th to the early 20th century.)

As well as rules governing the outward form of poetry, a system of poetic imagery already existed by this early period. The sequence of a poem, moreover, followed a fixed pattern (such as that for the *qaṣīdah*). Pre-Islāmic poetry was not written down but recited; and therefore sound and rhythm played an important part in its formation, and the *rāwīs* (reciters) were equally vital to its preservation. A *rāwī* was associated with some famous bard and, having learned his master's techniques, might afterward become a poet himself. This kind of apprenticeship to a master whose poetic style was thus continued became a common practice in the Muslim world (especially in Muslim India) right up to the 19th century.

The seven authors of *The Seven Odes*

From pre-Islāmic times the seven authors of *The Seven Odes*, already described, are usually singled out for special praise. Their poems and miscellaneous verses were collected during the 8th century and ever since have been the subject of numerous commentaries in the East. They have been studied in Europe since the early 19th century.

The poet Imru' al-Qays (died *c.* AD 550), of the tribe of Kindah, was foremost both in time and in poetic merit. He was a master of love poetry; his frank descriptions of dalliance with his mistresses are considered so seductive that (as orthodox Puritanism claims) the Prophet Muḥammad called him "the leader of poets on the way to Hell." His style is supple and picturesque. It grips the attention whether his poems sing of his love adventures or describe a seemingly endless rainy night. Of all classical Arabic poets he is probably the one who appeals most to modern taste. At the other extreme stands Zuhayr, praising the chiefs of the rival tribes of 'Abs and Dhubyān for ending a long feud. He is chiefly remembered for his serious *qaṣīdah* in which, old, wise, and experienced, he meditates upon the terrible escalation of war. Various aspects of Bedouin life, as well as the attitude of the Arabs to the rulers of the small kingdom of al-Ḥīrah on the Euphrates, are reflected in the poems of an-Nābighah adh-Dhubyānī, 'Amr, and Ṭarafah. The boastful pride of the self-centred Arab warrior can be observed best in the poems of al-Ḥārith, who became proverbial for his arrogance. 'Antarah, son of a black slave girl, won such fame on the battlefield and for his poetry that he later became the hero of an Arabic folk romance.

Two other masters can stand beside these seven. Exciting for their savagery and beauty are some poems by Ta'abbaṭa Sharran and Shanfarā, both outlaw warriors. Their verses reveal the wildness of Bedouin life, with its ideals of bravery, revenge, and hospitality. Ta'abbaṭa Sharran is the author of a widely translated "Song of Revenge" (for his uncle), composed in a short, sharp metre. Shanfarā's *lāmīyah* (literally "poem rhyming in l") vividly, succinctly, and with a wealth of detail tells of the experiences to be had from life in the desert. This latter poem has sometimes been considered a forgery, created by a learned grammarian. The suggestion highlights the question, often posed, of how much pre-Islāmic poetry is genuine and how much is the product of later scholars. Some modern critics—without proper justification—would dismiss the entire *corpus* as counterfeit.

Prose. While poetry forms the most important part of early Arabic literature and is an effective historical preservation of the Arabs' past glory, there is also a quantity of prose. Of special interest is the rhymed prose (*saj'*) peculiar to soothsayers, which developed into an important form of ornate prose writing in every Islāmic country. Tales about the adventures and battle days of the various tribes (*ayyām al-'Arab*, or "The Days of the Arabs") were told and handed down from generation to generation, usually interspersed with pieces of poetry. Proverbs and proverbial sayings were as common as in most cultures at a comparable level of development. The "literary" genre most typical of Bedouin life is the *musāmarah*, or "nighttime conversation," in which the central subject is elaborated not by plot but by carrying the listener's mind from topic to topic through verbal associations. Thus, the language as language played a most important role. The *musāmarah* form inspired the later *maqāmah* literature.

The *musāmarah*: "nighttime conversation"

It has been said—and this certainly holds true for the *musāmarah*—that Arabic literature demands attention from its listeners only in short bursts; for listeners are carried from verse to verse, from anecdote to anecdote, from pun to pun, along a theme whose broad outline is entirely familiar. Western Orientalists have for this reason spoken of the "molecular," or "atomic," structure both of classical Arabic literature and of traditional Islāmic thought. An audience listening to one of the ancient bards—or to a modern poet or orator in the Muslim world—would be able to listen without tiring. The sheer emotive power of the Arabic language to enrapture and bewitch its listeners by sound alone should be kept in mind when considering any piece of Arabic literature. Only a people endowed with peculiar sensibility to the word could properly appreciate the refinement of pre-Islāmic poetry and be ready to accept the concept of divine revelation appearing through the word in the Qur'ān.

EARLY ISLĀMIC LITERATURE

With the coming of Islām the attitude of the Arabs toward poetry seems to have changed. The new Muslims, despite their long-standing admiration for powerful language, often shunned poetry as reminiscent of pagan ideals now overthrown. For the Qur'ān, in *sūrah* 26:225 ff., condemned the poets "who err in every valley, and say what they do not do. Only the perverse follow them!" The Qur'ān, as the uncreated word of God, was now considered the supreme manifestation of literary beauty. It became the basis and touchstone of almost every cultural and literary activity and attained a unique position in Arabic literature.

Age of the caliphs. It might be expected that a new and vigorous religion would stimulate a new religious literature to sing of its greatness and glory. This, however, was not the case. Maybe the once boastful poets felt, at least for a while, that they were nothing but humble servants of Allāh. At any rate, no major poet was inspired by the birth and astonishingly rapid expansion of Islām. Only much later did poets claim that their work was the "heritage of prophecy" or draw upon a tradition that calls the tongues of the poets "the keys of the treasures beneath the Divine Throne." The old, traditional literary models were still faithfully followed: a famous ode by Ka'b, the son of Zuhayr, is different from pre-Islāmic poetry only insofar as it ends in praise of the Prophet, imploring his forgiveness, instead of eulogizing some Bedouin leader. Muḥammad's rather mediocre eulogist, Ḥassān ibn Thābit (died *c.* 659), also slavishly repeated the traditional patterns (even including the praise of wine that had been such a common feature of pre-Islāmic poetry at the court of al-Ḥīrah, despite the fact that wine had been by then religiously prohibited).

Influence of pre-Islāmic poetry

Religious themes are to be found in the *khuṭbah*s, or Friday sermons, which were delivered by governors of the provinces. In these *khuṭbah*s, however, political considerations frequently overshadow the religious and literary aspects. The *quṣṣāṣ* (storytellers), who interpreted verses from the Qur'ān, attracted large audiences and may be regarded as the inventors of a popular religious prose. Their interpretations were highly fanciful, however, and hardly squared with the theologian's orthodoxy.

The desire to preserve words of wisdom is best reflected in the sayings attributed to 'Alī, the fourth caliph (died 661). These, however, were written down, in superbly concise diction, only in the 10th century under the title *Nahj al-balāghah* ("The Road of Eloquence"), a work that is a masterpiece of the finest Arabic prose and that has inspired numerous commentaries and poetical variations in the various Islāmic languages.

Umayyad dynasty. The time of the "Four Righteous Caliphs," as it is called, ended with 'Alī's assassination in 661. The Umayyad dynasty then gained the throne, and a new impetus in poetry soon became perceptible. The Umayyads were by no means a pious dynasty, much enjoying the pleasures of life in their residence in Damascus and in their luxurious castles in the Syrian desert. One of their last rulers, the profligate al-Walīd ibn Yazīd (died 744), has become famous not so much as a conqueror (although in 711 the Muslims reached the lower Indus basin, Transoxania, and Spain) but as a poet who excelled in frivolous love verses and poetry in praise of wine. He was fond of short, light metres to match his subjects and rejected the heavier metres preferred by *qaṣīdah* writers. His verses convey a sense of ease and gracious living. Al-Walīd was not, however, the first to attempt this kind of poetry: a remarkable poet from Mecca, 'Umar ibn Abī Rabī'ah (died *c.* 712 or 720), had contributed in large measure to the separate development of the love poem (*ghazal*) from its subordinate place as the opening section of the *qaṣīdah*. Gentle and charming, in attractive and lively rhythms, his poems sing of amorous adventures with the ladies who came to Mecca on pilgrimage. His gay, melodious poems still appeal to modern readers.

The love poets of Medina

In Medina, on the other hand, idealized love poetry was the vogue; its invention is attributed to Jamīl (died 701), of the tribe ʿUdhrah, "whose members die when they love." The names of some of these "martyrs of love," together with the names of their beloved, were preserved and eventually became proverbial expressions of the tremendous force of true love. Such was Qays, who went mad because of his passion for Laylā and was afterward known as Majnūn (the "Demented One"). His story is cherished by later Persian, Turkish, and Urdu poets; as a symbol of complete surrender to the force of love, he is dear both to religious mystics and to secular poets.

Notwithstanding such new developments, the traditional *qaṣīdah* form of poetry was by no means neglected during the Umayyad period. Moreover, as the satirists of Iraq rose to fame, the *naqāʾiḍ* ("polemic poetry matches") between Jarīr (died *c.* 729) and al-Farazdaq (died *c.* 728 or 730) excited and delighted tribesmen of the rival settlements of Basra and Kūfah (places that later also became rival centres of philological and theological schools). The work of these two poets has furnished critics and historians with rich material for a study of the political and social situation in the early 9th century. The wealth of al-Farazdaq's vocabulary led one of the old Arabic critics to declare: "If Farazdaq's poetry did not exist, one-third of the Arabic language would be lost." Philologists, eager to preserve as much of the classical linguistic heritage as possible, have also paid a great deal of attention to the largely satirical poetry of al-Ḥuṭayʾah (died 674). The fact that Christians as well as Muslims were involved in composing classical Arabic poetry is proved by the case of al-Akhṭal (died *c.* 710), whose work preserves the pre-Islāmic tradition of al-Ḥīrah in authentic form. He is particularly noted for his wine songs. Christians and Jews had been included among the pre-Islāmic poets.

Prose literature was still restricted to religious writing. The traditions of the Prophet began to be compiled, and, after careful sifting, those regarded as trustworthy were preserved in six great collections during the late 9th century. Two of these—that of al-Bukhārī and that of Muslim ibn al-Ḥajjāj—were considered second only to the Qurʾan in religious importance. The first studies of religious law and legal problems, closely connected with the study of the Qurʾān, also belong to that period.

The ʿAbbāsids. It was not until the ʿAbbāsids assumed power in 750, settling in Baghdad, that the golden age of Arabic literature began. The influx of foreign elements added new colour to cultural and literary life. Hellenistic thought and the influence of the ancient cultures of the Near East, for example, contributed to the rapid intellectual growth of the Muslim community. Its members, seized with insatiable intellectual curiosity, began to adapt elements from all the earlier high cultures and to incorporate them into their own. They thus created the wonderful fabric of Islāmic culture that was so much admired in the Middle Ages by western Europe. Indian and Iranian threads were also woven into this fabric, and a new sensitivity to beauty in the field of poetry and the fine arts was cultivated.

The classical Bedouin style was still predominant in literature and was the major preoccupation of grammarians. These men were, as the modern critic Sir Hamilton Gibb has emphasized, the true humanists of Islām. Their efforts helped to standardize "High Arabic," giving it an unchangeable structure once and for all. By now the inhabitants of the growing towns in Iraq and Syria were beginning to express their love, hatred, religious fervour, and frivolity in a style more appealing to their fellow townsmen. Poets no longer belonged exclusively to what had been the Bedouin aristocracy. Artisans and freed slaves, of non-Arab origin, were included among their number. Bashshār ibn Burd (died *c.* 784), the son of a Persian slave, was the first representative of the new style. This ugly, blind workman excelled as a seductive love poet and also as a biting satirist—"Nobody could be secure from the itch of his tongue," it was later said—and he added a new degree of expressiveness to the old forms. The category of *zuhdīyāt* (didactic-ascetic poems) was invented by the poet Abū al-ʿAtāhīyah (died 825 or 826) from Basra, the centre of early ascetic movements. His pessimistic thoughts on the transitory nature of this world were uttered in an unpretentious kind of verse that rejected all current notions of style and technical finesse. He had turned to ascetic poetry after efforts at composing love songs.

The works of Abū Nuwās

The same is said of Abū Nuwās (died *c.* 813), the most outstanding of the ʿAbbāsid poets. His witty and cynical verses are addressed mainly to handsome boys; best known are his scintillating drinking songs. His line "Accumulate as many sins as you can" seems to have been his motto; and compared with some of his more lascivious lines, even the most daring passages of pre-Islāmic poetry sound chaste. Abū Nuwās had such an incomparable command over the language, however, that he came to be regarded as one of the greatest Arabic poets of all time. Nevertheless, orthodox Muslims would quote of him and of his imitators the Prophet's alleged saying that "poetry is what Satan has spit out," since he not only described subjects prohibited by religious law but praised them with carefree lightheartedness.

The "new" style. The new approach to poetry that developed during the 9th century was first accorded scholarly discussion in the *Kitāb al-badīʿ* ("Book of the Novel and Strange") by Ibn al-Muʿtazz (died 908), caliph for one day, who laid down rules for the use of metaphors, similes, and verbal puns. The ideal of these "modern" poets was the richest possible embellishment of verses by the use of tropes, brilliant figures of speech, and farfetched poetic conceits. Many later handbooks of poetics discussed these rules in minute detail, and eventually the increasing use of rhetorical devices no longer produced art but artificiality. (Ibn al-Muʿtazz was himself a fine poet whose descriptions of courtly life and nature are lovely; he even tried to compose a tiny epic poem, a genre otherwise unknown to the Arabs.) The "modern" poets, sensitive to colours, sounds, and shapes, also were fond of writing short poems on unlikely subjects: a well-bred hunting dog or an inkpot; delicious sweetmeats or jaundice; the ascetic who constantly weeps when he remembers his sins; the luxurious garden parties of the rich; an elegy for a cat; or a description of a green ewer. Their amusing approach, however, was sooner or later bound to lead to mannered compositions. The growing use of colour images may be credited to the increasing Persian influence upon ʿAbbāsid poetry; for the Persian poets were, as has been often observed, on the whole more disposed to visual than to acoustic imagery.

Tendency toward mannered writing

New attitudes toward love, too, were being gradually developed in poetry. Eventually, what was to become a classic theme, that of *ḥubb ʿudhrī* ("ʿUdhrah love")—the lover would rather die than achieve union with his beloved—was expounded by the Ẓāhirī theologian Ibn Dāʾūd (died 910) in his poetic anthology *Kitāb az-zahrah* ("Book of the Flower"). This theme was central to the *ghazal* poetry of the following centuries. Although at first completely secular, it was later taken over as a major concept in mystical love poetry. (The first examples of this adoption, in Iraq and Egypt, took place in Ibn Dāʾūd's lifetime.) The wish to die on the path that leads to the beloved became commonplace in Persian, Turkish, and Urdu poetry; and most romances in these languages end tragically. Ibn Dāʾūd's influence also spread to the western Islāmic world. A century after his death, the theologian Ibn Ḥazm (died 1064), drawing upon personal experiences, composed in Spain his famous work on "pure love" called *Ṭawq al-ḥamāmah* (*The Ring of the Dove*). Its lucid prose, interspersed with poetry, has many times been translated into Western languages.

The conflict between the traditional ideals of poetry and the "modern" school of the early ʿAbbāsid period also led to the growth of a literary criticism, the criteria of which were largely derived from the study of Greek philosophy.

Traditional poetry, meanwhile, was not neglected. But its style was somewhat modified in accordance with the new ideas. Two famous anthologies of Bedouin poetry, both called *Ḥamāsah* ("Poems of Bravery"), were collected by the Syrian Abū Tammām (died 845 or 846) and his disciple al-Buḥturī (died 897), both good classical poets in their own right. They provide an excellent survey of those poems from the stock of early Arabic poetry that were considered worth preserving. A century later Abū al-Faraj al-Iṣbahānī (died 967), in a multivolume work entitled *Kitāb al-aghānī* ("Book of Songs"), collected a great number of poems and biographical notes on poets and musicians. This material gives a colourful and valuable panorama of literary life in the first four centuries of Islām.

In the mid-10th century a new cultural centre emerged at the small court of the Ḥamdānids in Aleppo. Here the Central Asian scholar al-Fārābī (died 950) wrote his fundamental works on philosophy and musical theory. Here, too, for a while, lived Abū aṭ-Ṭayyib al-Mutanabbī (died 965), who is in the mainstream of classical *qaṣīdah* writers but who surpasses them all in the extravagance of what has been called his "reckless audacity of imagination." He combined some elements of Iraqi and Syrian stylistics with classical ingredients. His compositions—panegyrics of rulers and succinct verses (which are still quoted)—have never ceased to intoxicate the Arabs by their daring hyperbole, their marvelous sound effects, and their formal perfection. The Western reader is unlikely to derive as much aesthetic pleasure from Mutanabbī's poetry as does one whose mother tongue is Arabic. He will probably prefer the delicate verses about gardens and flowers by Mutanabbī's colleague in Aleppo, aṣ-Ṣanawbarī (died 945), a classic exponent of the descriptive style. This style in time reached Spain, where the superb garden and landscape poetry of Ibn Khafājah (died 1139) displayed an even higher degree of elegance and sensitivity than that of his Eastern predecessors.

Mutanabbī's poetry

Before turning to the development of prose, it is necessary to mention a figure unique among those writing in Arabic. This was Abū al-'Alā' al-Ma'arrī (died 1057), a blind poet of Syria, whose verses have appealed greatly to young Arabs of the present because of the poems' sincerity and humanity. But al-Ma'arrī's vocabulary is so difficult, his verses, with their double rhymes, are so compressed in meaning, that even his contemporaries, flocking to his lectures, had to ask him to interpret their significance. His outlook is deeply pessimistic and skeptical. Although his poems display a mastery of the Arabic traditional stylistic devices, they run counter to the conventional ideals of Arab heroism by speaking of bitter disappointment and emphasizing asceticism, compassion, and avoidance of procreation.

> Taking reason for his guide he judges men and things with a freedom which must have seemed scandalous to the rulers and privileged classes of the day. Among his meditations on the human tragedy a fierce hatred of injustice, hypocrisy, and superstition blazes out. Vice and folly are laid bare in order that virtue and wisdom may be sought . . .

says Reynold A. Nicholson, al-Ma'arrī's foremost interpreter in the West, who has also translated his *Risālat al-ghufrān* ("Epistle of Pardon"), which describes a visit to the Otherworld. Ma'arrī's extremely erudite book also contains sarcastic criticism of Arabic literature. His *Al-Fuṣūl wa al-ghāyāt* ("Paragraphs and Periods") is an ironic commentary on man and nature but is presented as a sequence of pious exhortations in rhymed prose. It has scandalized the pious, some of whom see it as a parody of the Qur'ān. Ma'arrī's true intention in this book, which came to light only recently, is unknown.

Development of literary prose. During the 'Abbāsid period, literary prose also began to develop. Ibn al-Muqaffa' (died *c.* 756), of Persian origin, translated the fables of Bidpai into Arabic under the title *Kalīlah wa Dimnah.* These fables provided Islāmic culture with a seemingly inexhaustible treasure of tales and parables, which are to be found in different guises throughout the whole of Muslim literature. He also introduced into Arabic the fictitious chronicles of the Persian *Khwatāy-nāmak* ("Book of Kings"). This was the source of a kind of pre-Islāmic mythology that the literati preferred above the somewhat meagre historical accounts of the Arab pagan past otherwise available to them. These activities demanded a smooth prose style, and Ibn al-Muqaffa' has therefore rightly been regarded as the inaugurator of what is called "secretarial literature" (that produced by secretaries in the official chancelleries). He also translated writings on ethics and the conduct of government, which helped to determine the rules of etiquette (*adab*). His works are the prototype of the "Mirror for Princes" literature, which flourished during the late Middle Ages both in Iran and in the West. In this literature, a legendary Persian counselor, Bozorgmehr, was presented as a paragon of wise conduct. Later, stories were invented that combined Qur'ānic heroes with historical characters from the Iranian past.

The work of al-Jāḥiẓ

A growing interest in things outside the limits of Bedouin life was reflected in a quantity of didactic yet entertaining prose by such masters as the broadminded and immensely learned al-Jāḥiẓ (died 869). In response to the wide-ranging curiosity of urban society, the list of his subjects includes treatises on theology, on misers, on donkeys, and on thieves. His masterpiece is *Kitāb al-Ḥayawān* ("Book of Animals"), which has little to do with zoology but is a mine of information about Arab proverbs, traditions, superstitions, and the like. Al-Jāḥiẓ's style is vigorous, loquacious, and uninhibited. His work, however, is not well constructed, and it lacks the clear sobriety of the "secretarial style." Yet the glimpses it affords into the life of various strata of society during the 9th century have rightly attracted the special interest of Western scholars. Less impressive, but almost as multifaceted, are the treatises of Ibn Abī ad-Dunyā (died 894).

The concept of *adab* was soon enlarged to include not only educational prose dealing with etiquette for all classes of people but belles lettres in general. The classic example of Arabic style for prose writers in this field, accepted as such for almost a millennium, is the writing of the Persian Ibn Qutaybah (died 889). His *'Uyūn al-akhbār* ("Fountains of Stories"), in 10 books, each dealing with a given subject, provided a model to which numberless essayists in the Muslim world conformed. In his book on poetry and poets, Ibn Qutaybah dared, for the first time, to doubt openly that pre-Islāmic poetry was incomparable. The most vigorous prose style was achieved by Abū Ḥayyān at-Tawḥīdī (died 1023), who portrayed the weaknesses of the two leading viziers, both notorious for their literary ambitions, ". . . with such bitterness," as Gibb remarks, "that the book was reputed to bring misfortune upon all who possessed a copy." This work, like others by Tawḥīdī that have quite recently been discovered, reveals the author's sagacity and striking eloquence. His correspondence on problems of philosophy with Miskawayh (died 1030), the author of a widely circulated book on ethics and of a general history, helps to complete the picture of this extraordinary writer.

Some time about 800 the Arabs had learned the art of papermaking from the Chinese. Henceforth, cheap writing material was available, and literary output was prodigious. The *Fihrist* ("Index"), compiled by the bookseller Ibn an-Nadīm in 988, gave a full account of the Arabic literature extant in the 10th century. This Index covered all kinds of literature, from philology to alchemy; but most of these works unfortunately have been lost. In those years manuals of composition (*inshā'*) were written elaborating the technique of secretarial correspondence, and they grew into an accepted genre in Arabic as well as in Persian and Turkish literature. The devices thought indispensable for elegance in modern poetry were applied to prose. The products were mannered, full of puns, verbal tricks, riddles, and the like. The new style, which was also to affect the historian's art in later times, makes a good deal of this post-classical Arabic prose look very different from the terse and direct expression characteristic of the early specimens. Rhymed prose, which at one time had been reserved for such religious occasions as the Friday sermons, was now regarded as an essential part of elegant style.

Rhetorical artistry in the *maqāmah*

This rhetorical artistry found its most superb expression in the *maqāmah,* a form invented by Badī' az-Zamān al-Hamadhānī (died 1008). Its master, however, was al-Ḥarīrī (died 1122), postmaster (head of the intelligence service) at Basra and an accomplished writer on grammatical subjects. His 50 *maqāmah*s, which tell the adventures of Abū Zayd as-Sarūjī, with a wealth of language and learning, come closer to the Western concept of short story than anything else in classical Arabic literature. They abound in verbal conceits, ambivalence, assonance, alliteration, palindromes; they change abruptly from earnest to jest, from the crude to the most sublime, as the modern scholar G.E. von Grunebaum has pointed out in his evaluation of this form, which he regards as the most typical literary reflection of the Islāmic spirit. The work of al-Ḥarīrī has certainly been widely admired in the East; it has been imitated in Syriac and in Hebrew and has formed part of the syllabus in Muslim high schools of India. The pleasure to be derived from the brilliant artifice and ingenuity behind such compositions has led to their being imitated in other literary fields: quite often, in later Persian literature, one finds poems—sometimes whole books—composed of letters without diacritical marks (which distinguish otherwise similar-looking letters) or even made up entirely of unconnected letters. Even a commentary on the Qurʾān, in undotted letters, has been written in India (by Fayẓī, died 1595).

Achievements in the western Muslim world. The Arabic literature of Moorish Spain and of the whole Maghrib developed parallel with that of the eastern countries but came to full flower somewhat later. Córdoba, the seat of the Umayyad rulers, was the centre of cultural life. Its wonderful mosque has inspired Muslim poets right up to the 20th century (such as Sir Muḥammad Iqbāl, whose Urdu ode, "The Mosque of Córdoba," was written in 1935). Moorish Spain was a favourite topic for reformist novelists of 19th-century Muslim India, who contrasted their own country's troubled state with the glory of classical Islāmic civilization. Moorish Spain reached its cultural, political, and literary heyday under ʿAbd ar-Raḥmān III (912–961). Literary stylistic changes, as noted in Iraq and Syria, spread to the west: there the old Bedouin style had always been rare and soon gave way to descriptive and love poetry. Ibn Hāniʾ (died 973) of Seville has been praised as the Western counterpart of al-Mutanabbī, largely because of his eulogies of the Fāṭimid caliph al-Muʿizz, who at that time still resided in North Africa. The entertaining prose style of Ibn ʿAbd Rabbihi (died 940) in his *al-ʿIqd al-farīd* ("The Unique Necklace") is similar to that of his elder contemporary Ibn Qutaybah, and his book in fact became more famous than that of his predecessor. Writers on music and philology also flourished in Spain; literary criticism was practiced by Ibn Rashīq (died 1064) and, later, by al-Qarṭājannī (died 1285) in Tunis. Ibn Ḥazm (died 1064), theologian and accomplished writer on pure love, has already been mentioned.

Writers on music and philology

Philosophy: Averroës and Avicenna. Philosophy, medicine, and theology, all of which flourished in the ʿAbbāsid East, were also of importance in the Maghrib; and from there strong influences reached medieval Europe. The influences often came through the mediation of the Jews, who, along with numerous Christians, were largely Arabized in their cultural and literary outlook. The eastern Muslim countries could boast of the first systematic writers in the field of philosophy, including al-Kindī (died *c.* 870), al-Fārābī (died 950), and especially Avicenna (Ibn Sīnā, died 1037). Avicenna's work in philosophy, science, and medicine was outstanding and was appreciated as such in Europe. He also composed religious treatises and tales with a mystical slant. One of his romances was reworked by the Maghribi philosopher Ibn Ṭufayl (died 1185) in his book *Ḥayy ibn Yaqẓān* ("Alive Son of Awake"), or *Philosophus Autodidactus* (the title of its first Latin translation, made in 1671). It is the story of a self-taught man who lived on a lonely island and who, in his maturity, attained the full knowledge taught by philosophers and prophets. This theme was elaborated often in later European literature.

The dominating figure in the kingdom of the Almohads, however, was the philosopher Averroës (Ibn Rushd, died 1198), court physician of the Berber kings in Marrākush (Marrakech) and famous as the great Arab commentator on Aristotle. The importance of his frequently misinterpreted philosophy in the formation of medieval Christian thought is well known. Among his many other writings, especially notable is his merciless reply to an attack on philosophy made by Ghazālī (died 1111). Ghazālī had called his attack *Tahāfut al-falāsifah* (*The Incoherence of the Philosophers*), while Averroës' equally famous reply was entitled *Tahāfut at-tahāfut* (*The Incoherence of the Incoherence*). The Persian-born Ghazālī had, after giving up a splendid scholarly career, become the most influential representative of moderate Ṣūfism. His chief work, *Iḥyāʾ ʿulūm ad-dīn* ("The Revival of the Religious Sciences"), was based on personal religious experiences and is a perfect introduction to the pious Muslim's way to God. It inspired much later religious poetry and prose. The numerous writings by mystics, who often expressed their wisdom in rather cryptic language (thereby contributing to the profundity of Arabic vocabulary), and the handbooks of religious teaching produced in eastern Arab and Persian areas (Sarrāj, Kalābādhī, Qushayrī, and, in Muslim India, al-Hujwīrī) are generally superior to those produced in western Muslim countries. Yet the greatest Islāmic theosophist of all, Ibn al-ʿArabī (died 1240), was Spanish in origin and was educated in the Spanish tradition. His writings, in both poetry and prose, shaped large parts of Islāmic thought during the following centuries. Much of the later literature of eastern Islām, particularly Persian and Indo-Persian mystical writings, indeed, can be understood only in the light of his teachings. Ibn al-ʿArabī's lyrics are typical *ghazal*s, sweet and flowing. From the late 9th century, Arabic-speaking mystics had been composing verses often meant to be sung in their meetings. At first a purely religious vocabulary was employed, but soon the expressions began to oscillate between worldly and heavenly love. The ambiguity thus achieved eventually became a characteristic feature of Persian and Turkish lyrics.

The pious Muslim's way to God

Among the Arabs, religious poetry mainly followed the classical *qaṣīdah* models, and the poets lavishly decorated their panegyrics to the Prophet Muḥammad with every conceivable rhetorical embellishment. Examples of this trend include *al-Burdah* ("The Mantle") of al-Buṣīrī (died 1298), upon which dozens of commentaries have been written (and which has been translated into most of the languages of Muslims because of the power to bless attributed to it). More sophisticated but less well known is an ode on the Prophet by the Iraqi poet Ṣafī ad-Dīn al-Ḥilli (died 1350), which contains 151 rhetorical figures. The "letters of spiritual guidance" developed by the mystics are worth mentioning as a literary genre. They have been popular everywhere; from the western Islāmic world the letters of Ibn ʿAbbād (died 1390) of Ronda (in Spain) are outstanding examples of this category, being written clearly and lucidly.

Geographical literature. The Maghrib also made a substantial contribution to geographical literature, a field eagerly cultivated by Arab scholars since the 9th century. The Sicilian geographer ash-Sharīf al-Idrīsī produced a famous map of the world and accompanied it with a detailed description in his *Kitāb nuzhat al-mushtāq fī ikhtirāq al-āfāq* ("The Delight of Him Who Wishes to Traverse the Regions of the World," 1154), which he dedicated to his patron, Roger II. The Spanish traveler Ibn Jubayr (died 1217), while on pilgrimage to Mecca, kept notes of his experiences and adventures. The resulting book became a model for the later pilgrims' manuals that are found everywhere in the Muslim world. The Maghribi explorer Ibn Baṭṭūṭah (died 1368/69 or 1377) described his extensive travels to the Far East, India, and the region of the Niger in a book filled with information about the cultural state of the Muslim world at that time. The value of his narrative is enhanced by the simple and pleasing style in which it is written.

Poetry. In the field of poetry, Spain, which produced a considerable number of masters in the established poetical forms, also began to popularize strophic poetry, possibly deriving from indigenous models. The *muwashshaḥ* ("girdled") poem, written in the classical short metres and arranged in four- to six-line stanzas, was elaborated, enriched by internal rhymes, and, embodying some popular expressions in the poem's final section, soon achieved a standardized form. The theme is almost always love. Among the greatest lyric poets of Spain was Ibn Zaydūn of Córdoba (died 1071), who was of noble birth. After composing some charming love songs dedicated to the Umayyad princess Wallādah, he turned his hand to po-

Spanish mastery of poetical forms

etic epistles. He is the author of a beautiful *muwashshaḥ* about his hometown, which many later poets imitated. When the *muwashshaḥ* was transplanted to the eastern Arabic countries, however, it lost its original spontaneity and became as stereotyped as every other lyric form of expression during the later Middle Ages. Another strophic form developed in Spain is the songlike *zajal* (melody), interesting for its embodiment of dialect phrases and the use of occasional words from Romance languages. Its master was Ibn Quzmān of Córdoba (died 1160), whose life-style was similar to that of Western troubadours. His approach to life as expressed in these melodious poems, together with their mixed idiom, suggests an interrelationship with the vernacular troubadour poetry of Spain and France.

Historiography: Ibn Khaldūn. Any survey of western Muslim literary achievements would be incomplete if it did not mention the most profound historiographer of the Islāmic world, the Tunisian Ibn Khaldūn (died 1406). History has been called the characteristic science of the Muslims because of the Qurʾānic admonition to discover signs of the divine in the fate of past peoples. Islāmic historiography has produced histories of the Muslim conquests, world histories, histories of dynasties, court annals, and biographical works classified by occupation—scholars, poets, and theologians. Yet, notwithstanding their learning, none of the earlier writers had attempted to produce a comprehensive view of history. Ibn Khaldūn, in the famous *Muqaddimah* or introduction to a projected general history, *Kitāb al-ʿibar,* sought to explain the basic factors in the historical development of the Islāmic countries. His own experiences, gained on a variety of political missions in North Africa, proved useful in establishing general principles that he could apply to the manifestations of Islāmic civilization. He created, in fact, the first "sociological" study of history, free from bias. Yet his book was little appreciated by his fellow historians, who still clung to the method of accumulating facts without shaping them properly into a well-structured whole. Ibn Khaldūn's work eventually attracted the interest of Western Orientalists, historians, and sociologists alike; and some of his analyses are still held in great esteem.

Significance of Ibn Khaldūn

Decline of the Arabic language. Ibn Khaldūn, who had served in his youth as ambassador to Pedro I the Cruel, of Castile, and in his old age as emissary to Timur, died in Cairo. After the fall of Baghdad in 1258, this city had become the centre of Muslim learning. Historians there recorded every detail of the daily life and the policies of the Mamlūk sultans; theologians and philologists worked under the patronage of Turkish and Circassian rulers who often did not speak a word of Arabic. The amusing, semicolloquial style of the historian Ibn Iyās (died after 1521) is an interesting example of the deterioration of the Arabic language. While classical Arabic was still the ideal of every literate man, it had become exclusively a "learned" language. Even some copyists who transcribed classical works showed a deplorable lack of grammatical knowledge. It is hardly surprising that poetry composed under such circumstances should be restricted to insipid versification and the repetition of well-worn clichés.

MIDDLE PERIOD: THE RISE OF PERSIAN AND TURKISH POETRY

The new Persian style. During the ʿAbbāsid period, the Persian influence upon the Arabic had grown considerably: at the same time, a distinct Modern Persian literature came into existence in northeastern Iran, where the house of the Sāmānids of Bukhara and Samarkand had revived the memory of Sāsānian glories.

The first famous representative of this new literature was the poet Rūdakī (died 940/941), of whose *qaṣīdah*s only a few have survived. He also worked on a Persian version of *Kalīlah wa Dimnah,* however, and on a version of the *Sendbād-nāmeh.* Rūdakī's poetry, modeled on the Arabic rules of prosody that without exception had been applied to Persian, already points ahead to many of the characteristic features of later Persian poetry. The imagery in particular is sophisticated, although when compared with the mannered writing of subsequent times his verse was considered sadly simple. From the 10th century onward, Persian poems were written at almost every court in the Iranian areas, sometimes in dialectical variants (for example, in Ṭabarestāni dialect at the Zeyārid court). In many cases the poets were bilingual, excelling in both Arabic and Persian (a gift shared by many non-Arab writers up to the 19th century).

Influence of Maḥmūd of Ghazna. The first important centre of Persian literature existed at Ghazna (present-day Ghaznī, Afg.), at the court of Maḥmūd of Ghazna (died 1030) and his successors, who eventually extended their empire to northwestern India. Himself an orthodox warrior, Maḥmūd in later love poetry was transformed into a symbol of "a slave of his slave" because of his love for a Turkmen officer, Ayāz. Under the Ghaznavids, lyric and epic poetry both developed, as did the panegyric. Classical Iranian topics became the themes of poetry, resulting in such diverse works as the love story of Vāmeq and ʿAzrā (possibly of Greek origin) and the *Shāh-nāmeh* ("Book of Kings"). A number of gifted poets praised Maḥmūd, his successors, and his ministers. Among them was Farrokhī of Seistan (died 1037), who was the author of a powerful elegy on Maḥmūd's death, one of the finest compositions of Persian court poetry.

Ferdowsī's literary achievement

Epic and romance. The main literary achievement of the Ghaznavid period, however, was that of Ferdowsī (died 1020). He compiled the inherited tales and legends about the Persian kings in one grand epic, the *Shāh-nāmeh,* which contains between 35,000 and 60,000 verses in short rhyming couplets. It deals with the history of Iran from its beginnings—that is, from the "time" of the mythical kings—passing on to historical events, giving information about the acceptance of the Zoroastrian faith, Alexander's invasion, and, eventually, the conquest of the country by the Arabs. A large part of the work centres on tales of the hero Rostam. These stories are essentially part of a different culture, thus revealing something about the Indo-European sources of Iranian mythology. The struggle between Iran and Tūrān (the central Asian steppes from which new waves of nomadic conquerors distributed Iran's urban culture) forms the central theme of the book; and the importance of the legitimate succession of kings, who are endowed with royal charisma, is reflected throughout the composition. The poem contains very few Arabic words and is often considered the masterpiece of Persian national literature, although it lacks proper historical perspective. Its episodes have been the inspiration of miniaturists since the 14th century. Numerous attempts have been made to emulate it in Iran, India, and Turkey.

Other epic poems, on a variety of subjects, were composed during the 11th century. The first example is Asadī's (died *c.* 1072) didactic *Garshāsb-nāmeh* ("Book of Garshāsb"), whose hero is very similar to Rostam. The tales of Alexander and his journeys through foreign lands were another favourite topic. Poetical romances were also being written at this time; they include the tale of *Varqeh o-Golshāh* by ʿEyyūqī (11th century) and *Vīs o-Rāmīn* by Fakhr od-Dīn Gorgānī (died after 1055), which has parallels with the Tristan story of medieval romance. These were soon superseded, however, by the great romantic epics of Neẓāmī of Ganja (died *c.* 1209), in Caucasia. The latter are known as the *Khamseh* ("Quintet") and, though the names of Vīs or Vāmeq continued for some time to serve as symbols of the longing lover, it was the poetical work of Neẓāmī that supplied subsequent writers with a rich store of images, similes, and stories to draw upon. The first work of his *Khamseh, Makhzan ol-asrār* ("Treasury of Mysteries"), is didactic in intention; the subjects of the following three poems are traditional love stories. The first is the Arabic romance of Majnūn, who went mad with love for Laylā. Second is the Persian historical tale of Shīrīn, a Christian princess, loved by both the Sāsānian ruler Khosrow II Parvīz and the stonecutter Farhād. The third story, *Haft peykar* ("Seven Beauties"), deals with the adventures of Bahrām Gūr, a Sāsānian prince, and seven princesses, each connected with one day of the week, one particular star, one colour, one perfume, and so on. The last part of the *Khamseh* is *Eskandar-nāmeh,* which relates the adventures of Alexander III the Great in Africa and Asia, as well as his discussions with the wise philosophers. It thus follows

The epics of Neẓāmī

the traditions about Alexander and his tutor, Aristotle, emphasizing the importance of a counselor-philosopher in the service of a mighty emperor. Neẓāmī's ability to present a picture of life through highly refined language and a wholly apt choice of images is quite extraordinary. Human feelings, as he describes them, are fully believable; and his characters are drawn with a keen insight into human nature. Not surprisingly, Neẓāmī's work inspired countless poets' imitations in different languages—including Turkish, Kurdish, and Urdu—while painters constantly illustrated his stories for centuries afterward.

Other poetic forms. In addition to epic poetry, the lesser forms, such as the *qaṣīdah* and *ghazal,* developed during the 11th and 12th centuries. Many poets wrote at the courts of the Seljuqs and also at the Ghaznavid court in Lahore, where the poet Masʿūd-e Saʿd-e Salmān (died 1121) composed a number of heartfelt *qaṣīdah*s during his political imprisonment. They are outstanding examples of the category of *ḥabsīyah* (prison poem), which usually reveals more of the author's personal feelings than other literary forms. Other famous examples of *ḥabsīyah*s include those written by the Arab knight Abū Firās (died 968) in a Byzantine prison; those by Muḥammad II al-Muʿtamid of Seville (died 1095) in the dungeons of the Almohads; those by the 12th-century Persian Khāqānī; those by the Urdu poets Ghālib, in the 19th, and Faiz, in the 20th century; and by the contemporary Turkish poet Nazim Hikmet (died 1963).

The most complicated forms were mastered by poets of the very early period, the limits of artificiality being reached in Azerbaijani *qaṣīdah*s by the poet Qaṭrān (died 1072), whose work displays virtuosity for virtuosity's sake. The court poets tried to top one another in the accumulation of complex metaphors and paradoxes, each hoping to win the coveted title "Prince of Poets." Anvarī (died *c.* 1189), whose patrons were the Seljuqs, is considered the most accomplished writer of panegyrics in the Persian tongue. His verses contain little descriptive material but abound in learned allusions. His "Tears of Khorāsān," mourning the passing of Seljuq glory, is among the best known of Persian *qaṣīdah*s. In the west of Iran, Anvarī's contemporary Khāqānī (died *c.* 1190), who wrote mainly at the court of the Shīrvān-Shāhs of Transcaucasia, is the outstanding master of the hyperbolic style. His mother was a Christian, and his imagery has more than the usual amount of allusions to Christian themes. His vocabulary seems inexhaustible; he uses uncommon rhetorical devices and very strong language. His poems, with their long chains of oath-formulae (*sowgandnāmeh*), are as impressive as his poignant antithetic formulations. Khāqānī's verses on the ruined Ṭāq Kisrā at Ctesiphon on the Tigris have become proverbial. His *qaṣīdah*s on the pilgrimage to Mecca, which also inspired his *maṡnavī, Tuḥfat al-ʿIrāqayn ol-ʿErāqeyn* ("Gift of the Two Iraqs"), translate most eloquently the feelings of a Muslim at the festive occasion. In the hand of lesser poets, however, *qaṣīdah* writing became more and more conventionalized, repeating outworn clichés and employing inflated terms entirely devoid of feeling.

Decline of the *qaṣīdah* form

Scholarship: al-Bīrūnī. The Ghaznavid and Seljuq periods produced first-rate scholars such as al-Bīrūnī (died 1048) who, writing in Arabic, investigated Hinduism and gave the first unprejudiced account of India—indeed, of any non-Islāmic culture. He also wrote notable books on chronology and history. In his search for pure knowledge he is undoubtedly one of the greatest minds in Islāmic history. Interest in philosophy is represented by Nāṣer-e Khosrow (died 1087/88) who acted for a time as a missionary for the Ismāʿīlī branch of Shīʿah Islām. His book about his journey to Egypt, entitled *Safar-nāmeh,* is a pleasing example of simple, clearly expressed, early Persian prose. His poetical works in the main seek to combine Greek wisdom and Islāmic thought: the gnostic Ismāʿīlī interpretation of Islām seemed, to him, an ideal vehicle for a renaissance of the basic Islāmic truths.

Robāʿīyāt: Omar Khayyam. The work done in mathematics by early Arabic scholars and by al-Bīrūnī was continued by Omar Khayyam (died 1122), to whom the Seljuq empire in fact owes the reform of its calendar. But Omar has become famous in the West through the free adaptations by Edward FitzGerald of his *robāʿīyāt.* These quatrains have been translated into almost every known language and are largely responsible for colouring European ideas about Persian poetry. The authenticity of these verses has often been questioned. The quatrain is an easy form to use—many have been scribbled on Persian pottery of the 13th century—and the same verse has been attributed to many different authors. The latest research into the question of the *robāʿīyāt* has established that a certain number of the quatrains can, indeed, be traced back to the great scientist who condensed in them his feelings and thoughts, his skepticism and love, in such an enthralling way that they appeal to every reader. The imagery he uses, however, is entirely inherited; none of it is original. (One of the most noted, and notorious, writers of this genre was the poetess Mahsaṭī [first half of the 12th century], who frequently addressed members of different professions in rather frivolous lines.) The quatrain was also popular as a means of embodying pieces of mystical wisdom. One has to do away with the old theory that the first author of such mystical *robāʿīyāt* was Abū Saʿīd ibn Abū al-Khayr (died 1049). A number of his contemporaries, however, including Bābā Ṭāher ʿOryān (died after 1055), used simpler forms of the quatrain, sometimes in order to express their mystical concepts.

The mystical poem. Whereas the mystical thought stemming from Iran had formerly been written in Arabic, writers from the 11th century onward turned to Persian. Along with works of pious edification and theoretical discussions, what was to be one of the most common types of Persian literature came into existence: the mystical poem. Khwajah ʿAbd Allāh al-Anṣārī of Herāt (died 1088), a prolific writer on religious topics in both Arabic and Persian, first popularized the literary "prayer," or mystical contemplation, written in Persian in rhyming prose interspersed with verses. Sanāʾī (died 1131?), at one time a court poet of the Ghaznavids, composed the first mystical epic, the didactic *Ḥadīqat al-ḥaqīqat wa sharīʿat aṭ-ṭarīqah* ("The Garden of Truth and the Law of the Path"), which has some 10,000 verses. In this lengthy and rather dry poem, the pattern for all later mystical *maṡnavī*s is established: wisdom is embodied in stories and anecdotes; parables and proverbs are woven into the texture of the story, eventually leading back to the main subject, although the argument is without thread and the narration puzzling to follow. Among Sanāʾī's smaller *maṡnavī*s, *Sayr al-ʿibād ilā al-maʿād* ("The Journey of the Servants to the Place of Return") deserves special mention. Its theme is the journey of the spirit through the spheres, a subject dear to the mystics and still employed in modern times as, for example, by Iqbāl in his Persian *Jāvīd-nāmeh* (1932). Sanāʾī's epic endeavours were continued by one of the most prolific writers in the Persian tongue, Farīd od-Dīn ʿAṭṭār (died *c.* 1220). He was a born storyteller, a fact that emerges from his lyrics but even more so from his works of edification. The most famous among his *maṡnavī*s is the *Manṭiq uṭ-ṭayr* (*The Conversation of the Birds*), modeled after some Arabic allegories. It is the story of 30 birds, who, in search of their spiritual king, journey through seven valleys. The poem is full of tales, some of which have been translated even into the most remote Islāmic languages. (The story of the pious Sheykh Ṣanʿān, who fell in love with a Christian maiden, is found, for example, in Kashmiri.) ʿAṭṭār's symbolism of the soul-bird was perfectly in accord with the existing body of imagery beloved of Persian poetry, but it was he who added a scene in which the birds eventually realize their own identity with God (because they, being *sī morgh,* or "30 birds," are identified with the mystical Sēmorgh, who represents God). Also notable are his *Elāhī-nāmeh,* an allegory of a king and his six sons, and his profound *Moṣībat-nāmeh* ("Book of Affliction"), which closes with its hero's being immersed in the ocean of his soul after wandering through the 40 stages of his search for God. The epic exteriorizes the mystic's experiences in the 40 days of seclusion.

Importance of Mawlānā Jalāl ad-Dīn ar-Rūmī. The most famous of the Persian mystical *maṡnavī*s is by Mawlānā ("Our Lord") Jalāl ad-Dīn ar-Rūmī (died 1273) and

Most famous Persian mystical *mašnavī*

is known simply as the *Mašnavī.* It comprises some 26,000 verses and is a complete—though quite disorganized—encyclopaedia of all the mystical thought, theories, and images known in the 13th century. It is regarded by most of the Persian-reading orders of Ṣūfis as second in importance only to the Qur'ān. Its translation into many Islāmic languages and the countless commentaries written on it up to the present day indicate its importance in the formation of Islāmic poetry and religious thought. Jalāl ad-Dīn, who hailed from Balkh and settled in Konya, the capital of the Rūm, or Anatolian Seljuqs (and hence was surnamed "Rūmī"), was also the author of love lyrics whose beauty surpasses even that of the tales in the *Mašnavī.* Mystical love poetry had been written since the days of Sanā'ī, and theories of love had been explained in the most subtle prose and sensitive verses by the Ṣūfis of the early 12th century. Yet Rūmī's experience of mystical love for the wandering mystic, Shams ad-Dīn of Tabriz, was so ardent and enraptured him to such an extent that he identified himself completely with Shams, going so far as to use the beloved's name as his own pen name. His dithyrambic lyrics, numbering more than 30,000 verses altogether, are not at all abstract or romantic. On the contrary, their vocabulary and imagery are taken directly from everyday life, so that they are vivid, fresh, and convincing. Often their rhythm invites the reader to partake in the mystical dance practiced by Rūmī's followers, the Mawlawīyah. His verses sometimes approach the form of popular folk poetry; indeed, Rūmī is reputed to have written mostly under inspiration; and despite his remarkable poetical technique, the sincerity of his love and longing is never overshadowed, nor is his personality veiled. In these respects he is unique in Persian literature.

Zenith of Islāmic literature. During the 13th century, the Islāmic lands were exposed, on the political plane, to the onslaught of the Mongols and the abolition of the 'Abbāsid caliphate, while vast areas were laid to waste. Yet this was in fact the period in which Islāmic literatures reached their zenith. Apart from Rūmī's superb poetry, written in the comparative safety of Konya, there was also the work of the Egyptian Ibn al-Fāriḍ (died 1235), who composed some magnificent, delicately written mystical poems in *qaṣīdah* style, and that of Ibn al-'Arabī, who composed love lyrics and numerous theosophical works that were to become standard. In Iran, one of the greatest literati, Moṣleḥ od-Dīn Sa'dī (died 1292), returned in about 1256 to his birthplace, Shīrāz, after years of journeying; his *Būstān* ("The Orchard") and *Golestān* ("Rose Garden") have been popular ever since. The *Būstān* is a didactic poem telling wise and uplifting moral tales, written in polished, easy-flowing style and a simple metre; the *Golestān,* completed one year later, in 1258, has been judged "... the finest flower that could blossom in a Sultan's garden" (Herder). Its eight chapters deal with different aspects of human life and behaviour. At first sight, its prose and poetical fragments appear to be simple and unassuming; but not a word could be changed without destroying the perfect harmony of the sound, imagery, and content. Sa'dī's *Golestān* is thus essential in discovering the nature of the finest Persian literary style. Since the mid-17th century, its moralizing stories have been translated into many Western languages. Sa'dī was likewise the author of some spirited *ghazal*s; he may have been the first writer in Iran to compose the sort of love poetry that is now thought of as characteristic of the *ghazal.* A few of his *qaṣīdah*s are also of note, although he is at his best in shorter forms. His elegant aphoristic poems, words of wisdom, and sensible advice all display what has been called the philosophy of common sense—how to act in any given situation so as to make the best of it both for oneself and others, basing one's conduct on the virtues of gentleness, elegance, modesty, and polite behaviour.

Sa'dī's "philosophy of common sense"

The influence of mysticism, on the one hand, and of the elaborate Persian poetical tradition, on the other, is apparent during the later decades of the 13th century, both in Anatolia and in Muslim India. The Persian mystic, Fakhr-ud-Dīn 'Irāqī (died 1289), a master of delightful love lyrics, lived for almost 25 years in Multān (in present-day Pakistan), where his lively *ghazal*s are still sung. His short treatises, in a mixture of poetry and prose (and written under Ibn al-'Arabī's influence), have been imitated often. While in Multān he may have met the young Amīr Khosrow of Delhi (died 1325), who was one of the most versatile authors to write in Persian, not only in India but in the entire realm of Persian culture. Amīr Khosrow, son of a Turkish officer, but whose mother was Indian, is often styled, because of the sweetness of his speech, "the parrot of India." (In Persian, it should be noted, parrots are always "sugar-talking"; they are, moreover, connected with Paradise and are thought of as wise birds—thus models of the sweet-voiced sage.) He wrote panegyrics of seven successive kings of Delhi and was also a pioneer of Indian Muslim music. Imitating Neẓāmī's *Khamseh,* Khosrow introduced a novelistic strain into the *mašnavī* by recounting certain events of his own time in poetical form, some parts of which are lyrics. His style of lyrical poetry has been described as "powdered"; and his *ghazal*s contain many of the elements that in the 16th and 17th centuries were to become characteristic of the "Indian" style. Khosrow's poetry surprises the reader in its use of unexpected forms and unusual images, complicated constructions and verbal plays, all handled fluently and presented in technically perfect language. His books on the art of letter writing prove his mastery of high-flown Persian prose. Khosrow's younger contemporary, Ḥasan of Delhi (died 1328), is less well known and had a more simple style. He nevertheless surpassed Khosrow in warmth and charm, qualities that have earned him the title of "the Sa'dī of Hindustan."

Turkish literature. As for the literary developments in Turkey around 1300, the mystical singer Yunus Emre is the first and most important in a long line of popular poets. Little is known about his life, which he probably spent not far from the Sakarya River of Asia Minor. Before him, in Central Asia, the religious leader Ahmed Yesevi (died 1166) had written some rather dry verses on wisdom in Turkish. Yunus, in Anatolia, however, was the first known poet to have caught something of Rūmī's fervour and translated it into a provincial setting, creating "... a Turkish vernacular poetry that was to be the model for all subsequent literary productions of popular religion." Sometimes he used the inherited Arabo-Persian prosody, but his best poems are those written in four-line verses using syllable-counting metres. Yunus drew heavily on the reservoir of imagery that had been collected by the great Persian writing mystics, notably Rūmī; but his classical technique did not hinder the expression of his own unself-conscious simplicity, which led him to introduce new images taken from everyday life in Anatolian villages. His *ilahi*s (hymns), probably written to be sung at the meetings of the Ṣūfis in the centres of their orders, are still loved by the Turks and memorized by their children.

Use of images from everyday life

Influence of Yunus Emre. The Turkish people rightly claim Yunus as the founder of Turkish literature proper. His poetry is considered the chief pillar of poetry of the Bektāshīyah Ṣūfī order, and many poets of this and other orders have imitated his style (though without reaching the same level of poetic truth and human warmth). Among the later poets claimed by the Bektāshīs may be mentioned Kaygusuz Abdal (15th century), who probably came from the European provinces of the Ottoman Empire. His verses are full of burlesque and even coarse images: in their odd mixture of worldliness and religious expression they are often as amusing as they are puzzling. In the 16th century, Pir Sultan Abdal (executed *c.* 1560) is noted for a few poems of austere melancholy. He was executed for collaboration with the Ṣafavids, the archenemies of the Ottomans; and in this connection it is worth remembering that the founder of the Iranian Ṣafavid dynasty, Shāh Esmā'īl I (died 1524), wrote Turkish poetry under the pen name Khaṭā'ī and is counted among the Bektāshī poets.

Religious poetry. Mystically tinged poetry has always been very popular in Turkey, both in cities and rural areas. The best loved religious poem of all was, and still is, Süleyman Çelebi's (died 1419) *Mevlûd,* a quite short *mašnavī* in honour of the Prophet Muḥammad's birth. This type of poetry has been known in the Islāmic countries since at least the 12th century and was soon adopted wherever Islām spread. There are a great number of *mevlûd* written in Turkish, but it was Süleyman Çelebi's unpretentious description of the great religious event that captured the hearts of the Turks; and it is still sung on many occasions (on the anniversary of a death, for example). The poem makes an excellent introduction to an understanding of the deep love for the Prophet felt by the pious Muslim.

Persian literature: 1300–1500. In the Iran of the Middle Ages, a vast number of poets flourished at the numerous courts. Not only professional poets but even the kings and princes contributed more or less successfully to the body of Persian poetry. Epics, panegyrics, and mystico-didactical poetry had all reached their finest hour by the end of the 13th century; the one genre to attain perfection slightly later was the *ghazal*, of which Moḥammad Shams od-Dīn Ḥāfeẓ (died 1389/90) is the incontestable master.

Lyric poetry: Moḥammad Shams od-Dīn Ḥāfeẓ. Ḥāfeẓ lived in Shīrāz; his pen name—"Who Knows the Qur-'ān by Heart"—indicates his wide religious education, but little is known about the details of his life. The same is true of many Persian lyrical poets, since their products rarely contain much trustworthy biographical material. Ḥāfeẓ's comparatively small collection of work—his *Dīvān* contains about 400 *ghazal*s—was soon acclaimed as the finest lyrical poetry ever written in Persian. The discussion of whether or not to interpret its wine and love songs on a mystical plane has continued for centuries. Yet this discussion seems sterile since Ḥāfeẓ, whose verbal images shine like jewels, is an outstanding exponent of the ambiguous and oscillating style that makes Persian poetry so attractive and so difficult to translate. The different levels of experience are all expressed through the same images and symbols: the beloved is always cruel, whether a chaste virgin (a rare case in Persian poetry!) or a professional courtesan, or, as in most cases, a handsome young boy, or God himself, mysterious and unattainable—or even, on the political plane, the remote despot, the wisdom of whose schemes must never be questioned by his subjects. Since mystical interpretation of the world order had become almost second nature to Persians during the 13th century, the human beloved could effortlessly be regarded as God's manifestation; the rose became a symbol of highest divine beauty and glory; the nightingale represented the yearning and complaining soul; wine, cup, and cupbearer became the embodiment of enrapturing divine love. The poets' multicoloured images were not merely decorative embroidery but were a structural part of their thought. One must not expect Ḥāfeẓ (or any other poet) to unveil his personal feelings in a lyrical poem of experience. But no other Persian poet has used such complex imagery on so many different levels with such harmonious and well-balanced lucidity as did Ḥāfeẓ. His true greatness lies in this rather than in the content of his poetry. It must be stressed again that, according to the traditional view, each verse of a *ghazal* should be unique, precious for its own sake, and that the apparent lack of logic behind the sequence of verses was considered a virtue rather than a defect. (It may help to think of the glass pieces in a kaleidoscope, which appear in different patterns from moment to moment, yet themselves form no logical pattern.) To what extent an "inner rhythm" and a "contrapuntal harmony" can be detected in Ḥāfeẓ's poetry is still a matter for discussion; but that he perfected the *ghazal* form is indisputable. Whether he is praised as a very human love poet, as an interpreter of esoteric lore, or, as has been recently suggested, as a political critic, his verses have a continuing appeal to all lovers of art and artistry.

Use of a standard set of images and symbols

Parodies of classic forms. Ḥāfeẓ's contemporary in Shīrāz was the satirist 'Obeyd-e Zākānī (died 1371), noted for his obscene verses (even the most moralistic and mystical poets sometimes produced surprisingly coarse and licentious lines) and for his short *maṡnavī* called *Mūsh o-gorbeh* ("Mouse and Cat"), an amusing political satire. Since few new forms or means of expression were open to them, 'Obeyd and other poets began ridiculing the classic models of literature: thus, Bosḥāq (died *c.* 1426) composed odes and *ghazal*s exclusively on the subject of food.

The Timurid period in Iran produced only moderately good poetry, despite the rulers' interest in art. Allegorical *maṡnavī*s were much in vogue, such as the *Shabestān-e khayāl* ("Bedchamber of Fantasy") by the prolific writer Fattāḥī of Nīshāpūr (died 1448) and *Gūy o-chowgān* ("Ball and Polo-stick") by 'Ārefī (died 1449); the latter work is an elaboration of the cliché that the lover is helpless before the will of his beloved, just as the ball is subject to the will of the polo-stick (". . . the head of the lover in the polo-stick of the beloved's tresses").

Poetry of the Timurid period

Eclecticism of 'Abd or-Raḥmān Jāmī. The last great centre of Islāmic art in the region of Iran was the Timurid court of Herāt, where Dowlatshāh (died 1494) composed his much-quoted biographical work on Persian poets. The leading figure in this circle was 'Abd or-Raḥmān Jāmī (died 1492), who is sometimes considered the last and most comprehensive of the "seven masters" in Persian literature, since he was a master of every literary genre and did not specialize in one form only, as Anvarī and Ḥāfeẓ, among others, had done. Jāmī wrote an excellent imitation of Neẓāmī's *Khamseh*, enlarging it by the addition of two mystical *maṡnavī*s into a septet called *Haft owrang* ("The Seven Thrones," or "Ursa Major"). His interest in Ṣūfism—he was initiated into the Naqshbandīyah order—is clear from his famous biographies of the Ṣūfī saints (which were an elaboration of a similar work by the 11th-century 'Abd Allāh al-Anṣārī). In imitation of Sa'dī, Jāmī also composed the *Bahārestān* ("Orchard of Spring"), written in prose interspersed with verses. He left no less than three large divans, which contain work of high quality and demonstrate his gift for inventing picturesque images. Although his work abounds in lavishly ornamented verses, his style on the whole lacks the perfect beauty of Ḥāfeẓ's lyrics and is already tending toward the heavier, more opaque "Indian" style. Jāmī also wrote treatises about literary riddles and various kinds of intellectual games, of which Muslim society in the late 15th century was very fond and which remain a feature of erudite Persian and Turkish poetry. His influence on the work of later poets, especially in Ottoman Turkey, was very powerful.

An interesting aspect of the Timurid court in Herāt was the attention given to Chagatai Turkish, which was spoken in the eastern regions of Islām. 'Alī Shīr Navā'ī, minister at the court (and a close friend of Jāmī), emphasized the beauties of his Turkic mother tongue as compared with Persian in his *Muḥākamat al-lughatayn* ("Judgment of the Two Languages"). He composed most of his lyrics and epics in Chagatai, which previously had been used by some members of the Timurid family and their courtiers for poetry but which became, thanks to him, an established literary medium. Even the arts-loving ruler of Herāt, Ḥusayn Bayqara (died 1506), wrote poetry in Turkic, following in every respect conventional literary taste.

Prose works: the "Mirror for Princes." During the first five centuries of Modern Persian literary life, a multitude of prose works were written. Among them, the "Mirror for Princes" deserves special mention. This genre, introduced from Persian into Arabic as early as the 8th century, flourished once more in Iran during the late 11th century. One important example is the *Qābūs-nāmeh* by the Zeyārid prince 'Onṣor ol-Ma'ālī Keykāvūs (died 1098), which presents "a miscellany of Islāmic culture in pre-Mongol times." At the same time, Niẓām al-Mulk (died 1092), the grand vizier of the Seljuqs, composed his *Seyāsat-nāmeh* ("Book of Government"), a good introduction to the statesman's craft according to medieval Islāmic standards. The *Seyāsat-nāmeh* was heavily influenced by pre-Islāmic Persian tradition. In the same period and environment, even a mystic like al-Ghazālī felt disposed to write a *Naṣīḥat al-mulūk* (Counsel for Kings), although the idealized relationship he makes between religious theory and practical statesmanship was not very realistic. A later mystic to compose a similar work was Sayyid 'Alī Hamadhānī (died 1385), who had settled in Kashmir and initiated its Sūfī poetry. Others, especially in India, exhorted rulers in their writings.

The "Book of Government"

Belles lettres. Belles lettres proper found a fertile soil in Iran. The fables of *Kalīlah wa Dimnah*, for example, were retold several times in Persian. The most famous version, though a rather turgid one, is called *Anvār-e soheylī* ("Lights of Canopus") and was composed by a famous mystic, Ḥoseyn Wāʿeẓ-e Kāshefī of Herāt (died 1504). The "cyclic story" form (in which several unconnected tales are held together by a common framework or narrator device), inherited from India, became as popular in Iran as it had been in the Arabic-speaking countries. The *Sendbād-nāmeh* and the *Ṭūṭī-nāmeh* ("Parrot Book"), which is based on Indian tales, are both good examples of the popular method whereby a variety of instructive stories are skillfully strung together within a basic "running" story. The first comprehensive collection of entertaining prose is *Jawāmiʿ al-ḥikayat* ("Collections of Stories"), a veritable storehouse of tales and anecdotes, by ʿOwfī (died *c.* 1230). Anecdotes were an important feature of the biographical literature that became popular in Iran and Muslim India. Biographies of the poets of a certain age or of a specified area were collected together. They provide the reader with few concrete facts about the subjects concerned; but they abound in anecdotes, sayings, and verses attributed to the subjects, thus preserving material that otherwise might have been lost. Many of these biographical manuals, such as ʿOwfī's *Lubāb al-albāb* ("Quintessence of the Hearts") or Dowlatshāh's *Tazkirat ash-shuʿarā* ("Biography of the Poets"), make agreeable reading. The authors concerned wished to demonstrate their own erudition and rhetorical technique as much as to immortalize their subjects; consequently, their books are important equally as stylistic documents and as historical sources. One of the most remarkable works in this field is *Chahār maqāleh* ("Four Treatises") by Neẓāmī-ye ʿArūẓī, a writer from eastern Iran. Written in about 1156, this little book is an excellent introduction to the ideals of Persian literature and its writers, discussing in detail what is required to make a perfect poet, giving a number of instances of the sort of poetic craftsmanship thought especially admirable, and allowing glimpses into the various arts in which the literary man was expected to excel.

"Anecdotal" writing

This tendency toward "anecdotal" writing, which is also manifest in the work of a number of Arab historians, can be observed in the cosmographical books and in some of the historical books produced in medieval Iran. Ḥamdollāh Mostowfī's (died after 1340) cosmography, *Nuzhat al-qulūb* ("Pleasure of the Hearts"), like many earlier works of this genre, underlined the mysterious aspects of the marvels of creation and was the most famous of several instructive collections of mixed folkloristic and scientific material. Early miniaturists, too, loved to illustrate the most unlikely tales and pieces of information given in such works. Historical writing proper had been begun by the Persians as early as the late 10th century, when Balʿamī's abridged translation of aṭ-Ṭabarī's (died 923) vast Arabic chronicle first acquainted them with this outstanding piece of early Arabic historical literature. The heyday of historiography in Iran, however, was the Il-Khanid period (mid-13th to mid-14th century). Iran was then ruled by the successors of Genghis Khan, and scholars began to extend their interest back to the history of pre-Islāmic Central Asia, whence the rulers had come. *Tārīkh-e jehān-goshāy* ("History of the World Conqueror") by ʿAṭā Malek-e Joveynī (died 1283) and *Jāmiʿ at-tawārīkh* ("Collector of Chronicles") by the physician and vizier Rashīd ad-Dīn (executed 1318) are both outstanding examples of histories filled with valuable information. Although the writing of history became a firmly established art in Iran and the adjacent Muslim countries, the facts were unfortunately all too often concealed in a bombastic style and a labyrinth of cumbersome, long-winded sentences. A history written by Vaṣṣāf (died 1323) is the most notorious example of turgidity, but even his style was surpassed by some later writers. These stylistic tendencies deeply influenced Turkish prose writing: 17th-century Turkish historical works, such as those of Peçevi (died *c.* 1650) and Naima (died 1716), for this reason almost defy translation. Later Persian prose in India suffered from the same defects. This development in Persian and Turkish prose is also reflected in the handbooks on style and letter writing that were written during the 14th and 15th centuries and afterward. They urged the practice of all the artificial tricks of rhetoric by this time considered essential for an elegant piece of prose.

Popular literature. Islāmic literatures, however, should not be thought to consist only of erudite and witty court poetry, of frivolous or melancholy love lyrics full of literary conceits, or of works deeply mystical in content. Such works are counterbalanced by a great quantity of popular literature, of which the most famous expression is *Alf laylah wa laylah* (*The Thousand and One Nights,* also known as *The Arabian Nights' Entertainment*). The tales collected under this title come from different cultural areas; their nucleus is of Indian origin, first translated into Persian as *Hazār afsānak* ("Thousand Tales") and then into Arabic. These fanciful fairy tales were later expanded with stories and anecdotes from Baghdad. Subsequently, some tales—mainly from the lower strata of society—about rogues, tricksters, and vagabonds were added in Egypt. Independent series of stories, such as that of Sindbad the Sailor, were also included. The entire collection is very important as a reflection of several aspects of Oriental folklore and allows, now and then, glimpses into the court life of the various dynasties. Since its first translation into French (1704), it has inspired many Western readers' dreams about the "romantic" East.

The Thousand and One Nights

From pre-Islāmic times the Arabs had recounted tales of the *ayyām al-ʿArab* ("Days of the Arabs"), which were stories of their tribal wars, and had dwelt upon tales of the heroic deeds of certain of their brave warriors, such as ʿAntarah. Modern research, however, suggests that his story in its present setting belongs to the period of the Crusades. The Egyptian queen, Shajar ad-Durr (died 1250), and the first brave Mamlūk ruler, Baybars I (died 1277), as well as the adventures of the Bedouin tribe Banū Hilāl on its way to Tunisia, are all the subjects of lengthy popular tales.

In Iran, many of the historical legends and myths had been borrowed and turned into high literature by Ferdowsī. Accounts of the glorious adventures of heroes from early Islāmic times were afterward retold throughout Iran, India, and Turkey. Thus, the *Dāstān-e Amīr Ḥamzeh,* a story of Muḥammad's uncle Ḥamzah ibn ʿAbd al-Muṭṭalib, was slowly enlarged by the addition of more and more fantastic details. This form of *dāstān,* as such literature is called, to some extent influenced the first attempts at novel writing in Muslim India during the 19th century. The epics of Köroğlu are common to both Iranian and Turkish tradition. He was a noble warrior-robber who became one of the central figures in folk literature from Central Asia to Anatolia.

Some popular epics were composed in the late Middle Ages, having as their basis local traditions. One such epic had as its basis the Turco-Iranian legend of an 8th-century hero, Abū Muslim, another the Turkish tales of the knight Dānishmend. Other epics, such as the traditional Turkish tale of Dede Korkut, were preserved by storytellers who improvised certain parts of their tales (which were noted down only afterward). Also, the role of the Ṣūfī orders and of the artisans' lodges in preserving and transmitting such semihistorical popular epics seems to have been considerable. Apart from heroic figures, the Muslim peoples further share a comic character—basically a type of low-class theologian, called Nasreddin Hoca in Turkish, Juḥā in Arabic, and Mushfiqī in Tadzhik. Anecdotes about this character, which embody the mixture of silliness and shrewdness displayed by this "type," have amused generations of Muslims.

Shortly after the introduction of the printing press, Turkey and Iran began to produce cheap books, sometimes illustrated, containing popular romantic love stories. Large numbers of fairy tales were published in these cheap editions, and still other fairy tales have been collected by European and Muslim folklorists.

Popularity of romantic love stories

A truly popular poetry is everywhere to be found: lullabies sung by Baluchi, Kurdish, and Ibo mothers have obvious similarities; workers sing little rhythmical poems to accompany their work, and nomads remember the adventures of their ancestors in their ballads. Such popular poems often contain dialect expressions, and the metres differ from the classical quantitative system. Some of these simple verses, such as a two-line *lanḍay* in Pashto, are among the most graceful products of Islāmic poetry. Many folksongs—lullabies, wedding songs, and dirges—

have a distinct mystical flavour and reflect the simple Muslim's love for the Prophet and his trust in God's grace even under the most difficult circumstances. Irony and wit are features of the riddle poem, a favourite form among Muslims everywhere. Folk poets were also fond of humorous descriptions of imaginary disputations between two entities—they might compose dialogues between coffee and tobacco (Morocco), between a big and a small mosque (Yemen), between a cat and a dog, or between a boy and a girl. This kind of literature in the semicolloquial or dialectical Arabic poetry of the 17th and 18th centuries in Yemen, Upper Egypt, and central Arabia would bear a thorough study. All the Iranian and Turkic languages, too, possess a rich heritage of popular poetry, which in many cases appeals more immediately to modern tastes than does the rather cerebral high literature of the urban and court cultures.

THE PERIOD FROM 1500 TO 1800

According to Persian tradition, the last classic author in literature was Jāmī, who died in 1492. In that year, Christopher Columbus discovered America, and the Christians reconquered Granada, the last Moorish stronghold of Spain. The beginning of the 16th century was as crucial in the history of the Muslim East as in that of the Western Hemisphere. In 1501, the young Esmāʿīl founded the Ṣafavid rule in Iran, and the Shīʿah persuasion of Islām was declared the state religion. At the same time, the kingdoms of the last Timurid rulers in Central Asia were overthrown by the Uzbeks, who, for a while, tried to continue the cultural tradition in both Persian and Turkic at their courts in Bukhara. In 1526, after long struggles, one member of the Timurid house, Bābur, laid the foundation of the Mughal Empire in India. In the Near East, the Ottoman Turks, having expanded their empire (beginning in the late 13th century) from northwestern Anatolia into the Balkans, conquered crumbling Mamlūk Egypt and adjacent countries, including the sacred places of Mecca and Medina in 1516–17. Thus, three main blocks emerged, and the two strongholds of Sunnī Islām—Ottoman Turkey and Mughal India—were separated by Shīʿah Iran.

Decentralization of Islāmic literatures. Ṣafavid Iran, as it happened, lost most of its artists and poets to the neighbouring countries: there were no great masters of poetry in Iran between the 16th and 18th centuries. And while the Persian Shāh Esmāʿīl wrote Turkish mystical verses, his contemporary and enemy, Sultan Selim I of Turkey (died 1520), composed quite elegant Persian *ghazal*s. Bābur (died 1530), in turn, composed his autobiography in Eastern Turkic.

Islāmic autobiographical literature

Bābur's autobiography is a fascinating piece of Turkish prose and at the same time one of the comparatively rare examples of Islāmic autobiographical literature. The classic example in this genre, however, was a lively Arabic autobiography by Usāmah ibn Munqidh (died 1188), which sheds much light upon the life and cultural background of a Syrian knight during the Crusades. A number of mystics, too, had written their spiritual autobiographies in a variety of languages, with varying degrees of artistic success. Bābur's book, however, gives a wonderful insight into the character of this intrepid conqueror. It reveals him as a master of concise, matter-of-fact prose, as a keen observer of daily life, full of pragmatic common sense, and also as a good judge of poetry. Bābur even went so far as to write a treatise in Turkish about versification. Many of his descendants, both male and female, inherited his literary taste and talent for poetry; among them are remarkably good poets in Persian, Turkish, and Urdu, as well as accomplished authors of autobiographies (Jahāngīr) and letters (Aurangzeb). Among the nobility of India, the Turkish language remained in use until the 19th century. Lovely Turkish verses were written, for example, by Akbar's general, Khān-e Khānān ʿAbd-ur-Raḥīm (died 1626), who was a great patron of fine arts and poetry.

In the Arab world, there was hardly a poet or original writer of note during the three centuries that followed the Ottoman conquest, apart from some theologians (ʿAbd al-Wahhāb ash-Shaʿrānī, died 1565; ʿAbd al-Ghanī an-Nābulusī, died 1731) and grammarians. Yet Arabic still remained the language of theology and scholarship throughout the Muslim world; both Turkey and India could boast a large number of scholars who excelled in the sacred language. In Ottoman Turkey, Taşköprüzâde (died 1560) compiled a historical survey of outstanding Turkish intellectuals in Arabic. Although a fine example of Islāmic learning, it does not compare in usefulness with the bibliographical work in Arabic by Hacı Halifa (Kâtib Çelebî; died 1658), which is a valuable source for modern knowledge of literary history.

New importance of Indian literature. India's share in the development of Arabic literature at this time was especially large. In addition to the quantity of theological work written in the language of the Qurʾān, from the conquest of Sind in 711 right up until the 19th century, much philosophical and biographical literature in Arabic was also being written in the subcontinent. Persian taste predominated in the northwest of India, but in the southern provinces there were long-standing commercial and cultural relationships with the Arabs, especially in Yemen and Ḥaḍramawt, and an inclination toward preserving these intact. Thus, much poetry in conventional Arabic style was written during the 16th and 17th centuries, mainly in the kingdom of Golconda. There are even attempts at the epic form. A century after the heyday of Arabic in the Deccan, Āzād Bilgrami (died 1786) composed numerous poetical and biographical works in Persian; but his chief fame was as the "Ḥassān of Hind," since he, like the Prophet Muḥammad's *protégé* Ḥassān ibn Thābit, wrote some powerful Arabic panegyrics in honour of the Prophet of Islām. He even attempted to make a comparison of the characteristics of Arabic and Sanskrit poetry and tried to prove that India was the real homeland of Islām. It should be added that al-Sayyid Murtaḍā az-Zabīd (died 1791), a leading philologist, author of the fundamental work of lexicography *Tāj al-ʿarūs* ("The Bride's Crown"), and commentator on Ghazālī's main work, was of Indian origin. Laudatory poems and belles lettres in Arabic were still popular in the early 19th century at the Shīʿite court of Lucknow, then the chief centre of Urdu poetry.

Poetry in conventional Arabic style

Indian literature in Persian. Nevertheless, the main contribution of Muslim India to high literature was made in the Persian tongue. Persian had been the official language of the country for many centuries. The numerous annals and chronicles that were compiled during the 14th and 15th centuries, as well as the court poetry, had been composed exclusively in this language even by Hindus. During the Mughal period, its importance was enhanced both by Akbar's attempt to have the main works of classical Sanskrit literature translated into Persian and by the constant influx of poets from Iran who came seeking their fortune at the lavish tables of the Indian Muslim grandees. At this time what is known as the "Indian" style of Persian emerged. The translations from Sanskrit enriched the Persian vocabulary, and new stories of Indian origin added to the reservoir of classical imagery. The poets, bound to the inherited genres of *maṡnavī, qaṣīdah,* and *ghazal,* tried to outdo each other in the use of complex rhyme patterns and unfamiliar, often stiff, metres. It became fashionable to conceive a poem according to a given *zamīn* ("ground"), in emulation of a classical model, and then to enrich it with newly invented tropes. The long-held ideal of "harmonious selection of images" was not always met. Difficult, even awkward grammatical constructions and inverted metaphors can be found. At times, pseudo-philosophical utterances in the second hemistich of a verse contrast strangely with semicolloquial expressions elsewhere. Objects recently introduced to India, such as the eyeglass or hourglass, were eagerly adopted as images by the poets, who wanted new-fangled conceits to bolster their tortuous inventiveness. Notwithstanding the colourful descriptive poems written in praise of such subjects as Mughal palaces, marvelously illuminated manuscripts, rare elephants, or court scenes, the general mood of lyric poetry became more gloomy. The transitory nature of the world, also a central theme in classical Persian poetry, was stressed and depicted in bizarre images: "burnt nest," "breakdown," "yawning" (indicating insatiable thirst); these were some of the new "stylish" words.

The works of 'Urfi

Yet some truly great poets are to be found even in this period. 'Urfī, who left Shīrāz for India and died in his mid-30s in Lahore (1592), is without doubt one of the few genuine masters of Persian poetry, especially in his *qaṣīdah*s. His verses pile up linguistic difficulties; yet their dark, glowing quality cannot fail to touch the hearts and minds even of critical modern readers—more so than the elegant but rather cerebral verses of his colleague Fayẕī (died 1595), one of Akbar's favourites. Fayẕī's brother Abū-ul-Faẕl 'Allāmī (died 1602), the author of an important, though biased, historical work, deeply influenced the Emperor's religious ideas. Among 17th-century Mughal court poets, the most outstanding is Abū Ṭālib Kalīm (died 1651), who came from Hamadan. Abounding in descriptive passages of great virtuosity, his poignant and often pessimistic verses have become proverbial, thanks to their compact diction and fluent style. Also of some importance is Ṣā'ib of Tabriz (died 1677), who spent only a few years in India before returning to Iran. Yet, of his immense poetical output (300,000 couplets), the great majority belongs to the stock-in-trade expression of the Persian-speaking world. Other poets described the lives and adventures of members of the royal families, usually in verbose *mas̄navī*s (this kind of descriptive historical poetry was practiced throughout Muslim India and also in Ottoman Turkey). Outside the Mughal environment, the lyrics and *mas̄navī*s by Ẓuhūrī (died 1615) at the court of Bijāpur are charming and enjoyable. The heir apparent of the Mughal Empire, Dārā Shikōh (executed 1659), also followed Akbar's path. His inclination to mysticism is reflected in both his prose and poetry. The Persian translation of the *Upaniṣad*s, which he sponsored (and in part wrote himself), enriched Persian religious prose and made a deep impression on European idealistic philosophy in the 19th century. A group of interesting poets gathered about him, none of them acceptable to orthodoxy. They included the convert Persian Jew Sarmad (executed 1661), author of mystical *robā'īyāt,* and the Hindu Brahman (died 1662), whose prose work *Chahār chaman* ("Four Meadows") gives an interesting insight into life at court. With the long rule of Dārā Shikōh's brother, the austere Aurangzeb (died 1707), the heyday of both poetry and historical writing in Muslim India was over. Once more, orthodox religious literature gained preeminence, while poets tried to escape into a fantasy world of dreams. The style of the two leading poets of this age, Nāṣir 'Alī Sirhindī (died 1697) and Mīrzā Bēdil (died 1721), is convoluted and obscure, prompting the Persian poet Ḥazīn (died 1766), who came to India in the early 18th century, to write ironic comments about its incomprehensibility. Bēdil, however, was a very interesting writer. His lyric poetry is difficult but often rewarding, while his many philosophical *mas̄navī*s deserve deep study. His prose work, interspersed with poetry, is called *Chahār 'unṣur* ("Four Elements") and contains some biographical details. His prose is nearly as difficult as his poetry, and consequently his works rarely have been read west of India. His poetry, however, has had a great influence in Afghanistan and Central Asia. Many Persian-speaking people there consider him the forerunner of Tadzhik literature, since virtually everyone in Bukhara and Transoxania who tried his hand at poetry followed Bēdil's example. His ideas, sometimes astoundingly modern and progressive, have also impressed the 20th-century poet and philosopher Muḥammad Iqbāl in Muslim India.

Bēdil's progressive ideas

With Bēdil, the "Indian summer" of Persian literature comes to an end, even though the output of Persian poetry and prose during the 18th century in the subcontinent was immense. Some of the biographical dictionaries and handbooks of mysticism are valuable for the scholar but are less interesting as part of the general history of literature. The main vehicle of poetry now became Urdu, while mystical poetry flourished in Sindhi and Punjabi.

Pashto poetry: Khushḥāl Khān Khaṭak. From the borderlands of the Persian-speaking zone, culturally under the Mughal rule, one man deserves special attention. The chief of the Pashtun tribe of Khaṭak, Khushḥāl Khān (died 1689), rightly deserves to be called the "father" of Pashto poetry, for he virtually created a literature of his own in his mother tongue. His skill in translating the sophisticated traditions of Persian literature into the not too highly developed idiom of the Pashtuns is astonishing. His lively lyric poems are his finest works, reflecting that passionate love of freedom for which he fought against the Mughals. The poems he wrote from prison in "hell-like hot India" are as dramatic as they are touching in their directness. Many members of his family took to poetry; and during the 18th century original works, both religious and secular, were composed in Pashto, and the classics of Persian literature were translated into that language.

Ottoman Turkey. The development of literature in Ottoman Turkey is almost parallel with that of Iran and India. Yunus Emre had introduced a popular form of mystical poetry; yet the mainstream of secular and religious literature followed Persian models (although it took some time to establish the Persian rules of prosody because of the entirely different structure of the Turkish language). In the religious field, the vigour and boldness expressed in the poems of Nesimî (executed 1417) left their traces in the work of later poets, none of whom, however, reached his loftiness and grandeur of expression. The 14th- and 15th-century representatives of the classical style had displayed great charm in their literary compositions, their verses simple and pleasing. Sultan Cem (Jem; died 1495), son of Mehmed the Conqueror, is an outstanding representative of their number. But soon the high-flown style of post-classical Persian was being imitated by Ottoman authors, rhetoric often being more important to them than poetical content. The work of Bâkî (Bāqī; died 1600) is representative of the entire range of these Baroque products. Yet his breathtaking command of language is undeniable; it is brilliantly displayed in his elegy on Süleyman the Magnificent. In his time, according to a popular saying, one could find "a poet under every stone of Istanbul's pavement." Istanbul was the unique cultural centre of the Near East, praised throughout the ages by all who lived in the imperial city.

Poetry of Fuzûlî of Baghdad. Much greater than most of these minor poets, however, was a writer living outside the capital, Fuzûlî of Baghdad (died 1556), who wrote in Arabic, Persian, and Azeri Turkish. Apart from his lyrics, his Turkish *mas̄navī* on the traditional subject of the lovers Majnūn and Laylā is admirable. From earliest times, Turkish poets had emulated the classical Persian romantic *mas̄navī*s, sometimes surpassing their models in expressiveness. Fuzûlî's diction is taut, his command of imagery masterly. His style unfortunately defies poetical translation, and his complicated fabric of plain and inverted images, of hidden and overt allusions is well-nigh impossible for all but the initiated Muslim reader to disentangle. Fuzûlî, moreover, like his fellow poets, would blend Arabic, Persian, and Turkish constructions and words to make up a multifaceted unit. The same difficulty is found in Turkish prose literature of the same period. It is a major task to unravel the long trailing sentences of a writer such as Evliya Çelebî (died after 1679), who, in an account of his travels (*Seyahatnâme*), has left extremely valuable information about the cultural climate in different parts of the Ottoman Empire.

Emulation of classical Persian *mas̄navī*s

Later developments. Growing interest in the Indo-Persian style, particularly in 'Urfī's *qaṣīdah*s, led the 17th-century Ottoman poets to a new integrated style and precision of diction. An outstanding representative was Nef'î, whose bent for merciless satire made him dreaded in the capital and eventually led to his assassination. At the start of the 18th century, a marked but short-lived movement in Turkish art known as the "Tulip Period" was the Ottoman counterpart of European Rococo. The musical poems and smooth *ghazal*s of Nedim (died 1730) reflect the manners and style of the slightly decadent, relaxed, and at times licentious high society of Istanbul and complement the miniatures of his contemporary Levnî. Good Turkish poetry is characterized by an easy grace, to be found even in such mystically tinged poems (thousands of which were written throughout the centuries) as those of Niyazî Misrî (died 1697). The Mevlevî (Mawlawī) poet Gâlib Dede (died 1799) was already standing at the threshold of what can now be recognized as modern poetical expression in some of the lyrical parts of his *maṡnavī,* called *Hüsn u aşk* ("Beauty and Love"), which brought fresh treatment to a well-worn subject of Iran's philosophical and secular literature. His work cannot be properly understood, however, without a thorough knowledge of mystical psychology, expressed in multivalent images.

Folk poetry. One branch of literature, however, was totally neglected by the sophisticated inhabitants of the Ottoman capital. Nobody thought much of the folk poets who wandered through the forgotten villages of Anatolia singing in simple syllable-counting verses of love, longing, and separation. The poems of the mid-17th-century figure Karacaoğlan, one of the few historically datable folk poets, give a vivid picture of village life, of the plight of girls and boys in remote Anatolian settlements. This kind of poetry was rediscovered only after the foundation of the Turkish Republic in 1923 and then became an important influence on modern lyric poetry.

EUROPEAN AND COLONIAL INFLUENCES: EMERGENCE OF WESTERN FORMS

The rise of nationalism. For the Islāmic countries, the 19th century marks the beginning of a new epoch. Napoleon's conquest of Egypt, as well as British colonialism, brought the Muslims into contact with a world whose technology was far in advance of their own. The West had experienced the ages of Renaissance, Reformation, and Enlightenment, whereas the once-flourishing Muslim civilization had for a long while been at a near stagnation point despite its remarkable artistic achievements. The introduction of Muslim intellectuals to Western literature and scholarship—the Egyptian aṭ-Ṭahṭāwī (died 1873), for example, studied in France—ushered in a new literary era the chief characteristic of which was to be "more matter, less art." The literatures from this time onward are far less "Islāmic" than those of the previous 1,000 years, but new intellectual experiences also led to "the liberation of the whole creative impulse within the Islāmic peoples" (Kritzeck). The introduction of the printing press and the expansion of newspapers helped to shape a new literary style, more in line with the requirements of the modern times, when "the patron prince has been replaced by a middle-class reading public" (Badawi). Translations from Western languages provided writers with the model examples of genres previously unknown to them, including the novel, the short story, and dramatic literature. Of those authors whose books were translated, Guy de Maupassant, Sir Walter Scott, and Anton Chekhov have been most influential in the development of the novel and the novella. Important also was the ideological platform derived from Tolstoy, whose criticism of Western Christianity was gratefully adopted by writers from Egypt to Muslim India. Western influences can further be observed in the gradual discarding of the time-hallowed static (and turgid) style of both poetry and prose; in the tendency toward simplification of diction; and in the adaptation of syntax and vocabulary to meet the technical demands of emulating Western models. Contact with the West also encouraged a tendency toward retrospection. Writers concentrated their attention on their own country and particular heritage, such as the "pharaoic myth" of Egypt, the Indo-European roots of Iran, and the Central Asian past of Turkey. In short, there was an emphasis on differentiation, inevitably leading to the rise of nationalism, instead of an emphasis on the unifying spirit and heritage of Islām.

Muslim intellectuals' introduction to Western scholarship

Arab literatures. Characteristically, therefore, given this situation, the heralds of Arab nationalism (as reflected in literature) were Christians. The historical novels of Jurjī Zaydān (died 1914), a Lebanese living in Egypt, made a deep impression on younger writers by glorifying the lion-hearted national heroes of past times. Henceforth, the historical novel was to be a favourite genre in all Islāmic countries, including Muslim India. The inherited tradition of the heroic or romantic epic and folktale was blended with novelistic techniques learned from Sir Walter Scott. Two writers in the front rank of Arab intellectuals were: Amīr Shakīb Arslān (died 1946), of Druze origin, and Muḥammed Kurd ʿAlī (died 1953), the founder of the Arab Academy of Damascus, each of whom, by encouraging a new degree of awareness, made an important contribution to the education of modern historians and men of letters. An inclination toward Romanticism can be detected in prose writing but not, surprisingly, in poetry; thus, the Egyptian al-Manfalūṭī (died 1924) poured out his feelings in a number of novels that touch on Islāmic as well as national issues.

Poetry. It is fair to say of this transition period that the poetry being written was not as interesting as the prose. The *qaṣīdah*s of the "Prince of Poets," Aḥmad Shawqī (died 1932), are for the most part ornate imitations of classical models. Even the "Poet of the Nile," Muḥammad Ḥāfiẓ Ibrahim (died 1932), who was more interested in the real problems of the day, was nonetheless content to follow conventional patterns. In his poems, Khalīl Muṭrān (died 1949) attempted to achieve a unity of structure hitherto almost unknown; and he also adopted a more subjective approach to expressive lyricism. Thus, he can be said to have inaugurated an era of "Romantic" poetry, staunchly defended by those men of letters who had come under English rather than French influence. These included the poet and essayist Ibrāhīm al-Māzinī (died 1949) and the prolific writer of poetry and prose ʿAbbās Maḥmūd al-ʿAqqād (died 1964).

Prose. A major contribution to the development of modern prose in the Arabic language was made by a number of writers born between 1889 and 1902. One of them, the "humanist" Taha Hussein, became well known in the West as a literary critic who attacked the historical authenticity of pre-Islāmic poetry and stressed the importance of Greek and Latin for the literatures of the modern Near East. He is also the author of a successful novel called *The Tree of Misery;* but his best creative writing is in his autobiographical notes, *al-Ayyām* ("The Days"), which describe in simple language the life of a blind Egyptian village boy. Taha Hussein's generation became more and more absorbed by the problems of the middle classes (to which most of them belonged), and this led them to realism in fiction. Some turned to fierce social criticism, depicting in their writings the dark side of everyday life in Egypt and elsewhere. The leading writer of this group is Maḥmūd Taymūr, who wrote short stories, a genre developed in Arabic by a Lebanese Christian who settled in the United States, the noted and versatile poet Khalil Gibran (Jibrān Khalīl Jibrān; died 1931). Muḥammad Ḥusayn Haykal (died 1956), a leading figure of Egyptian cultural and political life and the author of numerous historical studies, touched for the first time, in his novel *Zaynab* (1913), on the difficulties of Egyptian villagers. This subject quickly afterward became fashionable, although not all the writers had firsthand knowledge of the feelings and problems of the fellahin. The most fertile author of this group was al-ʿAqqād, who tirelessly produced biographies, literary criticism, and romantic poetry. To what extent the Islāmic reform movement led by Muḥammad ʿAbduh (died 1905) and his disciples, which centred on the journal *al-Manār* ("The Lighthouse"), has influenced present-day Arabic prose style cannot yet be ascertained. It has, however, been important in shaping the religious outlook of many authors writing in the 1920s and 1930s.

The diaspora. A considerable amount of Arabic literature has been produced by numerous writers who settled in non-Islāmic countries, especially in the United States and Brazil. Most of these writers came from Christian Lebanese families. A feeling of nostalgia often led them to form literary circles or launch magazines or newspapers. (The Arabic-language newspaper *al-Hudā* [or *Al-Hoda,* "The Guidance"], established in 1898, was published in New York City as *al-Hudā al-jadīdah* [*Al-Hoda Aljadidah,* or "The New Al-Hoda," or "The New Guidance"].) It was largely because of their work that the techniques of modern fiction and modern free verse entered Arabic literature and became a decisive factor in it.

One of the best known authors in this group was Amīn ar-Riḥānī (died 1941), whose descriptions of his journeys through the Arab world are informative and make agreeable reading. The fact that so many Lebanese emigrated to foreign countries led to the creation of a standard theme in Lebanese fiction: the emigrant who returns to his village.

The recurring theme in Lebanese fiction

Iraqi modern literature is best represented by "the poet of freedom" Ma'rūf ar-Ruṣāfī (died 1945), and Jamīl Sidqī az-Zahāwī (died 1936), whose satire "Rebellion in Hell" has incurred the wrath of the traditionalists.

Turkish literatures. The same changing attitude toward the function of literature and the same shift toward realism can be observed in Turkey. After 1839, Western ideas and forms were taken up by a group of modernists: Ziya Paşa (died 1880), the translator of Rousseau's *Émile* (which became a popular textbook for 19th-century Muslim intellectuals), was among the first to write in a less traditional idiom and to complain in his poetry—just as Ḥālī was to do in India a few years later—about the pitiable conditions of Muslims under the victorious Christians. Ziya Paşa, together with Şinasi (died 1871) and Namık Kemal (died 1888), founded an influential Turkish journal, *Tasvir-i Efkâr* ("Picture of Ideas"). The essential theme of the articles, novels, poems, and dramas composed by these authors is their fatherland (*vatan*), and they dared to advocate freedom of thought, democracy, and constitutionalism. Abdülhak Hâmid (died 1935), though considerably their junior, shared in their activities. In 1879 he published his epoch-making *Sahra* ("The Country"), a collection of ten Turkish poems that were the first to be composed in Western verse forms and style. Later, he turned to weird and often morbid subject matter in his poetic dramas. He, like his colleagues, had to endure political restrictions on writing, imposed as part of the harsh measures taken by Sultan Abdülhamid II against the least sign of liberal thought. Influenced by his work, later writers aimed to simplify literary language: Ziya Gökalp (died 1924) laid the philosophical foundations of Turkish nationalism; and Mehmed Emin, a fisherman's son, sang artless Turkish verses of his pride in being a Turk, throwing out the heavy rhetorical ballast of Arabo-Persian prosody and instead turning to the language of the people, unadulterated by any foreign vocabulary. The stirrings of social criticism could be discerned after 1907. Mehmed Akif (died 1936), in his masterly narrative poems, gave a vivid critical picture of conditions in Turkey before World War I. His powerful and dramatic style, though still expressed in traditional metres, is a testimony to his deep concern for the people's sorrows. It was he who composed the Turkish National Anthem after Mustafa Kemal Atatürk's victory; but soon afterward he left the country, disappointed with the religious policies of the Kemalists.

Atatürk's struggle for freedom also marks the real beginning of modern Turkish literature. The mainstream of novels, stories, and poems written during the 19th century had been replete with tears, world-weariness, and pessimism. But a postwar novel, *Ateşten gömlek* ("The Fire Shirt"), written by a woman, Halide Edib, reflected the brave new self-awareness of the Turkish nation. Some successful short stories about village life came from the pen of Ömer Seyfeddin (died 1920). The most gifted interpreter and harshest critic of Turkey's social structure was Sabaheddin Ali, who was murdered on his flight to Bulgaria in 1948. His major theme was the tragedy of the lower classes, and his writing is characterized by the same merciless realism that was later to be a feature of stories by many left-wing writers throughout the Islāmic world. The "great old man of Turkish prose," Yakup Kadri Karaosmanoğlu, displayed profound psychological insight, whether ironically describing the lascivious life in a Bektāshī centre or a stranger's tragedy in an Anatolian village. Most of the Turkish novelists of the 1920s and 1930s concentrated on the problems of becoming a modern nation, and in particular they reinterpreted the role of women in a liberated society.

The "great old man of Turkish prose"

Literary energies were set completely free when Atatürk introduced the Latin alphabet in 1928, hoping that his people would forget their Islāmic past along with the Arabic letters. From this time onward, especially after the language reform that was meant to rediscover the pre-Islāmic roots of the Turkish language, Turkish literature followed the pattern of Western literature in all major respects, though with local overtones. Poets experimented with new forms and new topics. They discovered the significance of the Anatolian village, neglected—even forgotten—during the Ottoman period. Freeing themselves from the traditional rules of Persian poetry, they adopted simpler forms from Europe. In some cases the skillful blending of inherited Ottoman grace and borrowed French lyricism produced outstandingly beautiful poems, such as those of Ahmed Haşim (died 1933) and of Yahya Kemal Beyatlı (died 1958), in which the twilight world of old Istanbul is mirrored in soft, evocative hues and melodious words. At the same time, the figure of Nazım Hikmet (died 1963) looms large in Turkish poetry. Expressing his progressive social attitude in truly poetical form, he used free rhythmical patterns quite brilliantly to enrapture his readers; his style, as well as his powerful, unforgettable images, has deeply influenced not only Turkish but also progressive Urdu and Persian poetry from the 1930s onward.

Persian literatures. In Iran, the situation to a certain extent resembled that in Turkey. While the last "classical" poet, Qā'ānī (died 1854), had been displaying the traditional glamorous artistry, his contemporary, the satirist Yaghmā (died 1859), had been using popular and comprehensible language to make coarse criticisms of contemporary society. As in the other Islāmic countries, a move toward simplicity is discernible during the last decades of the 19th century. The members of the polytechnic college Dār ol-Fonūn (founded 1851), led by its erudite principal Reẕā Qolī Khān Hedāyat, helped to shape the "new" style by making translations from European languages. Shāh Naṣer od-Dīn himself described his journeys to Europe in the late 1870s in a simple, unassuming style and in so doing set an example to future prose writers.

At the turn of the century, literature became for many younger writers an instrument of modernization and of revolution in the largest sense of the word. No longer did they want to complain, in inherited fixed forms, of some boy whose face was like the moon. Instead, the feelings and situation of women were stated and interpreted. Their oppression, their problems, and their grievances are a major theme of literature in this transition period of the first decades of the 20th century. The "King of Poets," Bahār (died 1951), who had been actively working before World War I for democracy, now devoted himself to a variety of cultural activities. But his poems, though highly classical in form, were of great influence; they dealt with contemporary events and appealed to a wide public.

One branch of modern Persian literature is closely connected with a group of Persian authors who lived in Berlin after World War I. There they established the Kaviani Press (named after a mythical blacksmith called Kaveh, who had saved the Iranian kingdom), and among the poems they printed were several by 'Aref Qazvīnī (died 1934), one of the first really modern writers. They also published the first short stories of Moḥammad 'Alī Jamālzādeh, whose outspoken social criticism and complete break with the traditional inflated and pompous prose style inaugurated a new era of modern Persian prose. Many young writers adopted this new form, among them Ṣādeq Hedāyat (died 1951), whose stories—written entirely in a direct, everyday language with a purity of expression that was an artistic achievement—have been translated into many languages. They reflect the sufferings of living individuals; instead of dealing in literary clichés, they describe the distress and anxiety of a hopeless youth. The influence of Franz Kafka (some of whose work Hedāyat translated) is perceptible in his writing, and he has a tendency toward psychological probing shared by many Persian writers.

Authors of the Kaviani Press

As in neighbouring countries, women played a considerable role in the development of modern Persian literature. The lyrics of Parvīn E'teṣāmī (died 1940) are regarded as near classics, despite a trace of sentimentality in their sympathetic treatment of the poor. Some Persian writers whose left-wing political ideas brought them into conflict with the government left for the Tadzhik S.S.R. Of these, the gifted poet Lāhūtī (died 1957) is their most important representative.

India: Urdu and Persian. Persian literature in the Indian subcontinent did not have such importance as in earlier centuries, for English replaced Persian as the official language in 1835. Nevertheless, there were some outstanding poets who excelled in Urdu. One of them was Mīrzā Asadullāh Khān Ghālib (died 1869), the undisputed master of Urdu lyrics. He regarded himself, however, as the leading authority on high Persian style and was an accomplished writer of Persian prose and poetry. But much more important was a later poet, Sir Muḥammad Iqbāl (died 1938), who chose Persian to convey his message not only to the peoples of Muslim India but also to Afghans and Persians. Reinterpreting many of the old mystical ideas in the light of modern teachings, he taught the quiescent Muslim peoples self-awareness, urging them to develop their personalities to achieve true individualism. His first *mas̄navī*, called *Asrār-e khudī* (1915; "Secrets of the Self"), deeply shocked all those who enjoyed the dreamlike sweetness of most traditional Persian poetry. One of his later Persian works, *Payām-e Mashriq* (1923; "Message of the East"), is an effective answer to Goethe's *West-östlicher Divan* (1819). In the *Jāvīd-nāmeh* (1932) he poetically elaborated the old topic of the "heavenly journey," discussing with the inhabitants of the spheres a variety of political, social, and religious problems. Iqbāl's approach is unique. Although he used the conventional literary forms and leaned heavily on the inspiration of Jalāl ad-Dīn ar-Rūmī, he must be considered one of the select few poets of modern Islām who, because of their honesty and their capacity for expressing their message in memorable poetic form, appeal to many readers outside the Muslim world.

THE MODERN PERIOD

The modern period of Islāmic literatures can be said to begin after World War II. The topics discussed before then still appeared, but outspoken social criticism became an even more important feature. Literature was no longer a leisurely pastime for members of the upper classes. Writers born in the villages and from non-privileged classes began to win literary fame through their firsthand knowledge of social problems. Many writers started their careers as journalists, developing a literary style that retained the immediacy of journalistic observation.

Prose. In Egypt, a great change in literary preoccupations came about after 1952. The name of Najīb Maḥfūẓ is of particular importance. He was at first a novelist mainly concerned with the lower middle classes (his outstanding work is a trilogy dealing with the life of a Cairo family); but afterward he turned to socially committed literature, using all the techniques of modern fiction—of which he is the undisputed master in Arabic. Yūsuf Idrīs deals first and foremost with the problems facing poor and destitute villagers, a subject also treated in Sharqāwī's novel *al-Arḍ* (*The Earth;* 1954). In Turkey, Yaşar Kemal's village story *İnce Memed* has won acclaim for its stark realism. The young left-wing writers in Iraq and Syria share the critical and aggressive attitudes of their contemporaries in Turkey and Egypt and are involved in every political issue. Most of them have responded to the works of Bertolt Brecht and Karl Marx. They are also quite familiar with at least the externals of modern psychology. Freudian influence—often in its crudest form—can be detected in many modern short stories or novels in the Islāmic countries; and it is often the prelude to coarse descriptions of sexuality, appealing to the lowest instincts of the reading public. In the Near and Middle East, the existentialist philosophy gained many followers who tried to reflect its interpretation of life in their literary works. In fact, almost every current of modern Western philosophy and psychology, every artistic trend and attitude, has been eagerly adopted—though often only half-understood—by young Arab, Turkish, or Persian writers. Some of them, nevertheless, have achieved interesting results from time to time: an example is Laylā Baʿlabakkī, whose semiautobiographical novel, *I Live,* is regarded as an outstanding literary achievement in Arabic.

Poetry. *Arabic.* The new attitudes that have informed literature are even more conspicuous in poetry than in prose. Arabic poetry has at last freed itself completely from the fetters of classical tradition. Both French and English influences helped to shape the new art. The danger is that Western fashions are imitated uncritically, just as Arabic, Persian, or Turkish models were slavishly followed in the past. T.S. Eliot's poetry and criticism were influential in dethroning the Romanticism that many poets had adopted earlier, in the 1920s and '30s. One of the first and most important attempts at creating a modern Arabic poetic diction was made in the late 1940s by the Iraqi poet and critic Nāzik al-Malāʾikah, whose poems, in free but rhyming verse, give substance to the shadow of her melancholia. Free rhythm and a colourful imagination distinguish the best poems of the younger Arabs: even when their poems do not succeed, their experimentation, their striving for sincerity, their burning quest for identity, their rebellion against social injustice can be readily perceived. Indeed, one of the most noticeable aspects of contemporary Arabic poetry is its political engagement, evident in the poems of Palestinian writers such as Maḥmūd Darwīsh, whose verses once more prove the strength, expressiveness, and vitality of the Arabic language. An Iraqi, ʿAbdul Wahhāb al-Bayātī, combines political engagement with lyrical mysticism. Others, without withdrawing into a world of uncommitted dreams, manage to create an atmosphere that breaks up the harsh light of reality into its colourful components. Poets like the Lebanese Adonis (ʿAlī Aḥmad Saʿīd) and Tawfīq aṣ-Ṣāʾigh, or the Egyptian dramatist Ṣalāḥ ʿAbd aṣ-Ṣabur, make use of traditional imagery in a new, sometimes esoteric, often fascinating and daring way.

New use of traditional imagery

Persian. Almost the same situation developed in Iran. One notable poet was Forugh Farrokhzād, who wrote powerful yet very feminine poetry. Her free verses, interpreting the insecurities of the age, are full of longing; though often bitter, they are yet truly poetic. Poems by such critically minded writers as Seyāvūsh Kasrāʾī also borrow the classical heritage of poetic imagery, transforming it into expressions that win a response from modern readers.

Turkish. In Turkey, the adoption of Western forms began in the 1920s. Of major importance in modern Turkish literature was Orhan Veli Kanık, who combined perfect technique with "Istanbulian" charm. His work is sometimes melancholy, sometimes frivolous, but always convincing. He strongly influenced a group of poets whose names are connected with the avant-garde literary magazine *Varlik* ("Existence"). The powerful poetry of the leftist writer Nazım Hikmet has influenced progressive poets all over the Muslim world. It is still too early, however, to determine what will be most representative of modern Turkish poetry: a return to Anatolian subjects, sometimes in picturesque diction, influenced by earlier folk-poetry, or the continuation of lovely poems in praise of Istanbul; surrealism or a somewhat detached and ironic approach to a subject. The same question, indeed, could be raised of almost any contemporary literature in Islāmic lands.

Contemporary features. In the Arab-speaking world, the problem of language has loomed large for many years. Classical high Arabic is still the common literary language of Morocco and Iraq, Tunisia and Kuwait. Spoken Arabic in dialectal variations is beginning to be used—but tentatively—in higher literature. It is more frequently employed in the popular spheres of theatre and cinema. But the local differences that exist in Arabic spoken from country to country have today become perceptible in literature; popular grammatical forms and syntactical constructions are occasionally used in modern poetry. A special problem arises in the North African countries, where French continues to be the chief literary language for most writers, especially in Morocco and Algeria. Yet there is no hard and fast rule: a leading member of the Senegal community, Amadou Bamba, who founded the politically important group of the Murīdīs, wrote (quite apart from practical words of wisdom in his mother tongue) some 20,000 mystically tinged verses in classical Arabic.

The continuing problem of language

Throughout the Islāmic countries, the press and radio have helped to disseminate literary works; prizes for literary achievements have stimulated interest in writing; low-priced books have made the more or less valuable output of a growing number of writers available to the majority—the more so since literacy among the population steadily increases. But to what degree this means a continuation of the cultural role that Islāmic literatures have played in the formation and education of society over the centuries is not yet clear. Literature was never restricted to a privileged high society; in olden times even the illiterate villager and the "uneducated" womenfolk had a fund of poems, proverbs, songs, and quotations from classical sources that they knew by heart and to which they turned for both pleasure and spiritual strength.

One final symptom should be noted. The introduction of modern methods of criticism, of psychology and philosophy, has kindled a new interest in significant figures of the Islāmic past. Thus, to quote one instance, the figure of al-Ḥallāj (executed 922), who often served as a symbol figure of "the martyr of love" in both classical and folk poetry after the 11th century, has in recent years been made the subject of a Turkish drama, a Persian passion play, and an Arabic tragedy and plays an important role in Arabic, Turkish, Persian, and Indian Muslim lyrical poetry. He is interpreted as a symbol of suffering for one's ideals, and he is therefore acceptable both to conservative Muslims and to progressive social critics.

STUDY AND EVALUATION

Early Islāmic criticism. The development of literature during the early Middle Ages soon produced among the Arabs much lively literary criticism. Even the choice of quotations made by the ancient grammarians from the classical stock of poetry implies a degree of critical (though subjective) activity. Attempts toward making a more objective study of poetic technique were first made in the late 9th century, when for the first time "beauties" and "faults" of verses were discussed and the ideals of the "new style" were defined by Ibn al-Muʿtazz in his *Kitāb al-badīʿ*. The relation between *lafẓ* (word) and *maʿnā* (meaning) has been a matter of some controversy—many earlier critics stress the importance of outward form rather than of content. There was some question, too, as to whether the most "poetical" verse was that which was the most "untrue"—that is to say, hyperbolic—or that which was closer to the heart of things. The matter was debated along with the problem of inspiration and imagination and their function in poetry. The most thorough analysis of the art of poetry was made by ʿAbd al-Qāhir al-Jurjānī, who allowed equal weight to the idea and to the way it was expressed. An illuminating work about poetics was composed by the Tunisian critic al-Qarṭājannī (13th century), and this has been carefully studied by the German scholar Wolfhart Heinrichs in *Arabische Dichtung und griechische Poetik* (1969). This study analyzes al-Qarṭājannī's theories in relation to Aristotle's theories of poetics. (Heinrichs, one of the few Islāmic scholars specializing in the study of literary problems, has shown that classical Arabic criticism rarely interested itself in the poem as a whole but concentrated upon individual verses.) In later centuries, manuals of poetics and rhetoric written in every Islāmic country reveal the prevailing interest in purely formal problems.

The controversy over word and meaning

Modern criticism. A similar interest long dominated the work of Western Orientalists. The first scholars who attempted to introduce Persian poetry to Western readers (such as Sir William Jones in the 18th century) thought it necessary to compare it with the compositions of Greek and Latin poetry. The verbal ingenuity of Ḥarīrī's *Maqāmāt* attracted the European scholars, who took great pleasure in disentangling the grammatically difficult forms. Pre-Islāmic poetry at first interested only the grammarian-antiquarian until its importance as a source of knowledge of early Bedouin life was recognized. The art of versification and problems of classical Arabic metrics are forever matters of discussion among Orientalists.

Although a large amount of translation, mainly from Persian poetry, was produced in the 18th and 19th centuries, most of it suffered for lack of proper understanding: the translators took the poetical statements about wine and love or the outbursts against established religious forms at their face value and failed to recognize them for the stereotyped forms and images they are. A deep study of the imagery of Persian, Turkish, and Arabic is required before their poetry and belles lettres can be properly understood and enjoyed. This was realized as early as 1818 by the Austrian Orientalist Joseph von Hammer-Purgstall (whose own translations from the three great Islāmic languages are, nevertheless, failures).

In the 20th century the critical study of imagery in Oriental poetry was taken up by Hellmut Ritter in his booklet *Über die Bildersprache Niẓāmīs* (1927; "On the Imagery of Neẓāmī"), which gives a most sensitive philosophical interpretation of Neẓāmī's metaphorical language and of the role of imagery in the structure of Neẓāmī's thought. Ritter's criticism is basic to the study of many other Persian poets. Slightly later, the Polish scholar Tadeusz Kowalski tried to interpret the "molecular" structure of Arabic literature—the absence of large units of thought or architectural structure—typical of the greater part of Islāmic literatures, which might be described as "carpet-like." This "molecular" structure can be related to the atomist theories and occasionalist world view embodied in Islāmic theology, which, unlike Christianity, does not admit of secondary causes and requires only short spans of hope from the faithful. In a number of articles, and in many books, E.G. von Grunebaum has pioneered this interpretation of literary structure. Other important critical works include S.A. Bonebakker's book on the rhetorical importance of *tawrīyah* (ambiguous wording); Manfred Ullmann's excellent study of *rajaz*-poetry and its place in Arabic literature; and C.H. de Fouchécour's detailed analysis of the descriptions of nature in early Persian poetry.

Among the Arabs themselves, modern literary criticism began during the early 1920s. Most famous was Ṭaha Ḥussein's attempt to prove the whole corpus of pre-Islāmic poetry as counterfeit. All the Islāmic countries, from Turkey to Pakistan, and especially Iran have sponsored reviews in which Western-trained scholars critically survey the literary achievements of the Islāmic world.

A full evaluation of literature as the most faithful mirror of past (and to some extent present-day) Islāmic life is still lacking. Notwithstanding the conventionalized style of most Islāmic poetry, a deeper study of individual poets' expressions, use of verbal and nominal forms, rhythmical preferences, and the like would certainly reveal more about the personalities of outstanding writers. The impact of poetry on the Islāmic mind was, and to some extent still is, much deeper than a modern Western reader might suppose. The poets must be viewed, therefore, in relation to their society, for their work corresponded to the measure of receptiveness, their new modes of expression developed according to the widening awareness of their audiences. They had to use a language and imagery to which those whom they addressed were accustomed. A new idea, embodied in traditional imagery and a beguiling metre, could capture the attention of thousands of people. The role of the poet as religious and political herald (even though his political thought was all too often subservient to courtly flattery) was widely acknowledged, and the impact of a poet like Muḥammad Iqbāl bears witness to the real power of poetical expression. Thus, even the most conventional Persian or Turkish poem can reflect certain attitudes of the Muslim mind more accurately than many a learned lecture. A modern short story, even if not particularly well wrought, often tells the reader more about the feelings and reactions of the people than scholarly sociological research papers can. The magic of language is still a living force in the East.

(An.Sc.)

Music

The period of Islāmic music begins with the advent of Islām in about 610. A new art emerged, elaborated both from pre-Islāmic Arabian music and from important contributions by Persians, Byzantines, Turks, Berbers, and Moors. In this development the Arabian element acted as a catalyst, and, within a century, the new art was firmly established from Central Asia to the Atlantic. Such a fusion of musical styles succeeded because there were strong affinities between Arabian music and the music of the nations occupied by the expanding Arabic peoples. Not all Arab-dominated areas adopted the new art; Indonesia and parts of Africa, for example, retained native musical styles. The folk music of the Berbers in North Africa, the Moors in Mauretania, and other ethnic groups (*e.g.,* in Turkey) also remained alien to classical Islāmic music. The farther one looks from the axis reaching from the Nile Valley to Persia, the less one finds undiluted Islāmic music.

(It should be remembered that the word *music* and its concept were reserved for secular art music; separate names and concepts belonged to folk songs and to religious chants.)

NATURE AND ELEMENTS OF ISLĀMIC MUSIC

Islāmic music is characterized by a highly subtle organization of melody and rhythm, in which the vocal component predominates over the instrumental. It is based on the skill of the individual artist, who is both composer and performer and who benefits from a relatively high degree of artistic freedom. The artist is permitted, and indeed encouraged, to improvise. He generally concentrates on the details forming a work, being less concerned with following a preconceived plan than with allowing the music's structure to emerge empirically from its details. Melodies are organized in terms of *maqāmāt* (singular *maqām*), or "modes," characteristic melodic patterns with prescribed scales, preferential notes, typical melodic and rhythmic formulas, variety of intonations, and other conventional devices. The performer improvises within the framework of the *maqām*, which is also imbued with ethos (Arabic *ta'thīr*), a specific emotional or philosophical meaning attached to a musical mode. Rhythms are organized into rhythmic modes, or *īqā'āt* (singular *īqā'*), cyclical patterns of strong and weak beats.

Classical Islāmic music is the aristocratic music of the court and the upper class, which underwent development and modification in the hands of gifted musicians throughout several centuries. Rhythmic and melodic modes grew in number and complexity, and new vocal and instrumental genres arose. In addition, a body of theoretical works grew up, influencing both Islāmic and—in some cases—European music. Its later popularization did not alter its intimate and entertaining character.

The relation of music to poetry and dance. In pre-Islāmic times music was closely connected with poetry and dance. Being essentially vocal, pre-Islāmic music was an emotional extension of the solemn declamation of poems in Bedouin society. Later, the art of vocal composition itself was largely based upon prosody: only by respecting the poetic metre in the music could the text, when sung, be clear in meaning and correct in pronunciation and grammatical inflection. In turn, prosody itself was used to explain the musical rhythm.

The poet-musician

Words and rhetorical speech were the principal means through which the Bedouin expressed feelings. The *shā'ir*, or poet-musician, said to be possessed by supernatural powers, was feared and respected. His satirical song poems were a formidable arm against enemies, and his poems of praise enhanced the prestige of his tribe. Musician-poets, especially women, accompanied the warriors, inciting them by their songs, and those who fell in battle benefited from the elegies of the singer-poets. Musically, these elegies resembled the *ḥudā'* ("caravan song"), possibly used by camel drivers as a charm against the desert spirits, or *jinn*.

Music and dance were closely associated from early times. Bedouin music had a pronounced collective character, with well-defined functions and usages, and dance occupied an important place in Bedouin life. Most common was a simple communal dance that emphasized common, or social, rather than individual movement. Places of entertainment in the towns and oases employed professional dancers, mainly women. Art dancing embellished events in the courts of the Sāsānians, the pre-Islāmic rulers of Persia. In the Islāmic period, solo and ensemble forms of dance were an integral part of the intense musical activity in the palaces of the caliphs and in wealthy houses. Dance also was prominent in the *dhikr* ceremony of certain mystical fraternities; forms ranged from obsessional physical movements to refined styles similar to those of secular art dancing.

After the advent of Islām a deep change occurred in the social function of music. Emphasis was laid on music as entertainment and sensual pleasure rather than as a source of high spiritual emotion, a change mainly resulting from Persian influence. Knowledge of music was obligatory for the cultured person. Skilled professional musicians were highly paid and were admitted to the caliphs' palaces as courtesans and trusted companions. The term *ṭarab*, which designates a whole scale of emotions, characterizes the musical conception of the time and even came to mean music itself.

Music and religion. Fashionable secular music—and its clear association with erotic dance and drinking—stimulated hostile reactions from religious authorities. As Muslim doctrine does not sanction permitting or prohibiting a given practice by personal decision, the antagonists relied on forced interpretations of a few unclear passages in the Qur'ān (the sacred scripture of Islām) or on the *Ḥadīth* (traditions of the Prophet, sayings and practices that had acquired force of law). Thus both supporters and adversaries of music found arguments for their theses.

In the controversy, four main groups emerged: (1) uncompromising purists opposed to any musical expression; (2) religious authorities admitting only the cantillation of the Qur'ān and the call to prayer, or *adhān;* (3) scholars and musicians favouring music, believing there to be no musical difference between secular and religious music; and (4) important mystical fraternities, for whom music and dance were a means toward unity with God.

Religious music

Except in the Ṣūfī brotherhoods, Muslim religious music is relatively curtailed because of the opposition of religious leaders. It falls into two categories: the call to prayer, or *adhān* (in some places, *azān*), by the *mu'adhdhin*, or muezzin, and the cantillation of the Qur'ān. Both developed from relatively solemn cantillation to a variety of forms, both simple and highly florid. The cantillation of the Qur'ān reflected the ancient Arabic practice of declamation of poetry, with careful regard to word accents and inflections and to the clarity of the text. Yet it was possibly also influenced by early secular art song. Opponents of music considered the cantillation of the Qur'ān to be technically distinct from singing, and it acquired a separate terminology. Synagogues and the Eastern Christian churches, unhampered by such opposition, developed extensive musical repertories based on melodic modes: the Eastern churches used the eight modes of Byzantine music, while synagogue music followed the *maqām* system of Muslim art music.

Aesthetic traditions. Even in its most complicated aspects, Islāmic music is traditional and is transmitted orally. A rudimentary notational system did exist but it was used only for pedagogical purposes. A large body of medieval writing about music survives in which musical theory is related to various areas of intellectual activity, hence the extreme importance of understanding music as an element of the culture involved. The medieval writings fall mainly into two categories: (1) literary, encyclopaedic, and anecdotal sources, and (2) theoretical, speculative sources. The first group includes precious information on musical life, musicians, aesthetic controversies, education, and the theory of musical practice. The second deals with acoustics, intervals (distances between notes), musical genres, scales, measures of instruments, the theory of composition, rhythm, and the mathematical aspects of music. These documents show that, as in the modern era, medieval Islāmic music was principally an individual, soloistic art. Small ensembles were actually groups of soloists with the principal member, usually the singer, predominating. Being an essentially vocal music, it displayed many singing and vocal techniques, such as special vocal colour, guttural nasality, vibrato, and other stylistic ornaments. Although the music was based upon strict rules, preexisting melodies, and stylistic requirements, the performer enjoyed great creative freedom. The artist was expected to bring his contribution to a given traditional piece through improvisation, original ornamentation, and his own approach to tempo, rhythmic pattern, and the distribution of the text over the melody. Thus the artist functioned as both performer and composer.

The soloist's role

Melodic organization. Islāmic music is monophonic; *i.e.*, it consists of a single line of melody. In performance everything is related to the refinement of the melodic line and the complexity of rhythm. The notion of harmony is completely absent, although occasionally a simple combination of notes, octaves, fifths, and fourths, usually below the melody notes, may be used as an ornamentation. Among the elements contributing to the enrichment of the melody are microtonality (the use of intervals smaller than a Western half step or lying between a half step and a Western whole step) and the variety of intervals used. Thus the three-quarter tone, introduced into Islāmic music in the 9th or 10th century, exists alongside larger and smaller intervals. Musicians show a keen sensibility to nuances of pitch, often slightly varying even the perfect consonances, the fourth and fifth.

As the fourth is the basic melodic frame, theorists organized the intervals and their nuances into genres, or small units, often tetrachords (units the highest and lowest notes of which are a fourth apart), combining genres into larger units, or systems. More than 130 systems resulted; on these are based the musical scales of the *maqāmāt*, or modes. The scale of a *maqām* can thus be broken down into small units that are of importance in the formation of melodies. A *maqām* is a complex musical entity given distinct musical character by its given scale, small units, range and compass, predominant notes, and preexisting typical melodic and rhythmic formulas. It serves the musician as rough material for his own composition. Each *maqām* has a proper name that may refer to a place (as Hejaz, Iraq), to a famous man, or to an object, feeling, quality, or special event. Emotional or philosophical meaning (ethos, or *ta'thīr*) and cosmological background are attached to a *maqām* and also to the rhythmic modes. The Arabic term *maqām* is the equivalent of *dastgāh* in Persia, *naghmah* in Egypt, and *cbāṭ* in North Africa.

Rhythmic organization. Rhythms and their organization into cycles of beats and pauses of varying lengths (rhythmic modes, or *īqā'āt*) are much discussed in theoretical writings and are of supreme importance in performance. Each cycle consists of a fixed number of time units with a characteristic distribution of strong and weak beats and pauses. In performance some of the pauses may be filled in, but the underlying pattern must be maintained. Parallel to the growth of the number of melodic modes—from 12 in the 8th century to more than 100 in the 20th—is the increase in the number of rhythmic modes from eight in the 9th century to more than 100 in the 20th.

Musical forms. The repertoire in common use comprises a wide variety of forms. One category includes unmeasured improvised pieces, such as the *layālī*, in which the singer puts forth the characteristics of the *maqām*, using long vocalises and meaningless syllables. An equivalent instrumental improvisation is called *taqsīm*, and this in some cases may be accompanied by a uniform pulsation, called *taqsīm 'ala al-wuḥdah*. The category of metrical songs embraces various poetic forms and metric structures, such as *qaṣīdah*, *dor*, and *muwashshaḥ*. Both categories, metrical and unmeasured, are almost always accompanied by either one or more instruments to enrich the performance. Important traditional forms combined both categories to create large compositions similar to a suite, using vocal and instrumental features. The whole was linked by the unity of the mode and a defined rhythmical development. Examples are the Andalusian *nūbah*, which survives in North Africa, the Persian *dastgāh*, the Turkish *fāṣil*, the Egyptian *waṣla*, and the Iraqi *macam*. Under the pressure of modernization and westernization have emerged new forms showing the influence of light dance music, operetta, and musical comedy.

Instruments of music. Instrumental music is not considered an independent art from vocal music. Yet many instruments were fully described by early writers, and their use in folk, art, religious, and military music pointed out. The most favoured instrument of ancient Near Eastern civilization, the harp, was gradually overshadowed by both long- and short-necked lutes.

Percussion instruments. Among idiophones (instruments the hard bodies of which vibrate to produce sound) commonly used are the *qaḍīb* ("percussion stick"), the *zil* and *sunūj* ("cymbals"), and the *kāṣāt*, or small finger cymbals. Membranophones, or vibrating membrane instruments, include a variety of tambourines, or frame drums, which all fall under the generic name *duff*. These include the North African *ghirbāl* and *bendīr*, instruments that have a number of "snares" across the skin and are used for folk dances; and the *dā'irah*, or *ṭar*, with jingling plates or rings set in the frame. The *dā'irah* and the vase-shaped drum *darabukka* (in Iran, *ẓarb*) are used in folk and art music, and the small kettledrums *naqqārah* and *nuqayrat* are used in art music and in military music (such as janissary music, the Turkish ensemble adopted by European military musicians). The large two-headed cylindrical drum, the *ṭabl* (Turkish *davul*), is generally played with the oboe-like *zornā* or *gayta* in processions and open-air ceremonies.

Wind instruments. Classed with the *zornā* and *gayta* as aerophones, or wind instruments, are the *būq*, or horn, the *nafīr*, or long trumpet, and a variety of flutes called *nāy* or *shabbābah*. Clarinetlike (single-reed) double-piped instruments such as the *dunay*, *zammārah*, and *urghūl* are used in folk events and open-air ceremonies.

Stringed instruments. Chordophones, or stringed instruments, constitute the most important family. The favourite instrument of Islāmic classical music is the *'ūd*, a short-necked lute having four or five strings and resembling the Western lute, which derived from the *'ūd*. In addition to holding musical supremacy, it was important in medieval theoretical and cosmological speculations. It has two derivatives in North Africa, the *kuwītra* and the *gunbrī*. The long-necked lutes favoured in Turkey, Iran, and the countries eastward include the *ṭunbūr*, *tār*, and *setār*. Another plucked instrument is the *qānūn*, or trapezoid-shaped psaltery, played at least from early medieval times. The trapezoidal dulcimer, or *sanṭūr*, the strings of which are struck with two thin sticks, is widespread and is especially prominent in Persian art music. Bowed lutes, or fiddles, include the *rabāb*, used by epic singers and beggars, and the *kamān*, or *kamanjā*, a hemispherically-shaped fiddle the body of which, like that of the *rabāb*, is pierced by the length of wood forming the neck (such instruments are known as spike fiddles). The violin, played either on the knee like the *kamanjā*, or beneath the collarbone, is also common.

The relation of Islāmic music to music of other cultures. The relation of Islāmic music to the West reveals itself in both musical theory and practice. By the 9th century many Greek treatises had been translated into Arabic. Arabic culture preserved Greek musical writings, and most of those that reached the West did so in their Arabic versions. Arab theorists followed Greek models, often developing them further. The Muslim occupation of Spain and Portugal and the Crusades to the Near East brought Europeans in contact with Arabic theoretical writings and the flourishing Islāmic art music. Musical instruments such as the lute, the rebec (a small bowed instrument derived from the *rabāb*), and the kettledrum (in the form of a pair of small kettledrums called nakers, from the Arabic *naqqārah*) became firmly established in European music. Arabic writings were translated, among them the *De scientiis*, a work on the arts and sciences by the great 10th-century philosopher and musician al-Fārābī (Latinized as Alpharabius). Such translations give further indication of the influence exerted by Muslim writers. Arabian influence on European medieval music is difficult to prove. Borrowed elements were possibly completely transformed. The influence of Islāmic music on European music is, at present, a subject of controversy.

Medieval contacts with Europe

As early as 711, Arab conquerors reached India, and Mongol and Turkmen armies later invaded the Near East, with resulting contact between Islāmic and Far Eastern music. There are similarities between the modal systems of India (the *rāgas*) and of the Near East (the *maqām* system) and between some cosmological and ethical conceptions of music. The migration of musical instruments from the Islāmic area to the Far East can also be traced. The Chinese oboe, the *sona*, apparently derived its name from its Near Eastern counterpart, the *zornā*, or *sornā*. The Indian long-necked lute sitar, having a different number of strings from the Persian *setār*, received its name, and perhaps part of its form, from the *setār*. The Chinese dulcimer, *yang ch'in* ("foreign zither"), originated in the Middle Eastern *sanṭūr*. On the other hand, the musical instruments appearing in the pre-Islāmic Ṭāq-e Bostān reliefs in Persia show a mouth organ similar to the Chinese *sheng*, indigenous to the Far East.

THE HISTORY OF ISLĀMIC MUSIC

The earliest extant writings on Islāmic music are from the end of the 9th century, more than 250 years after the advent of Islām. In the absence of historical documents, musicians, writers, and philosophers began to speculate on the origins of their music. They filled the gaps by legendary sources or vague traditions. Thus Lamak is said to have made the first lute from the leg of his dead son, whose loss he lamented with it. His lamentation is considered to be the first song.

The pre-Islāmic period. In nomadic encampments music emphasized every event in man's life, embellished social meetings, incited the warriors, encouraged the desert traveler, and exhorted the pilgrims to the black stone of the Kaʿbah (in Mecca), a holy shrine even in pre-Islāmic times. Among the earliest songs were the *ḥudāʾ* from which the *ghināʾ* derived, the *naṣb, sanad, rukbānī,* and the *hazāj,* a dancing song. In the markets of the Arabs, particularly the fair at the western Arabian town of ʿUkāẓ, competitions of poetry and musical performances were held periodically, attracting the most distinguished poet-musicians. Their music, more sophisticated than that practiced in the nomadic encampments, was related to that of the *qaynāt* ("singing girls"), who performed at court, in noble households, and in scattered taverns. Cultural contact with Byzantium was strong in the kingdom of Ghassān, where, in the 7th century, five Byzantine *qaynāt* were known to have performed songs of their homeland at court. The culture of the other Arabic kingdom of al-Ḥīrah under the Lakhmid dynasty was closely connected with that of Persia under the pre-Islāmic Sāsānian empire. The Sāsānians esteemed both secular and religious music. In the belief of the Mazdak sect (a dualistic Persian religion related to Manichaeanism, a Gnostic religion), music was considered as one of the four spiritual powers. In the king's entourage musicians occupied high rank. Some became famous, such as Bārbad, to whom is attributed the invention of the complicated pre-Islāmic system of modes. The compositions of Bārbad, who became a model of artistic achievement in Arabic literature, survived at least until the 10th century.

Courtly music

The beginning of Islām and the first four caliphs. Muḥammad was said to have been hostile to music and musicians; yet there are indications that he tolerated functional music such as war songs, pilgrimage chants, and public or private festival songs. In addition, he himself instituted in 622 or 623 the *adhān* ("call to prayer"), chanted by the *muʾadhdhin* (muezzin). For this task he chose the Abyssinian singer Bilāl, who became the patron of the *muʾadhdhin* and their guilds throughout the Islāmic world. Within 12 years after Muḥammad's death, the armies of Islām took possession of Syria, Iraq, Persia, Armenia, Egypt, and Cyrenaica (in modern Libya). The contact with the refined cultures of the conquered and the appearance of a new class of warriors who benefited from the spoils of the conquered nations deeply affected Arabian society. In spite of the austere regime of the four orthodox caliphs (632–660), joy of life and eagerness for pleasure dominated the two holy cities of Mecca and Medina. Wealthy men acquired slave musicians, who were often liberated and became the pillars of musical life. The wealthy competed with one another in the brilliance of the concerts held in their houses, and in sophisticated literary and musical salons, contests revealed and rewarded the best talents. In this milieu the great Islāmic musical tradition began to take shape, to be firmly established and codified in subsequent periods. A new generation of musicians was educated in the traditional manner and refined through constant hearing of the best music performed by the best masters. Through the contributions of the conquered "foreigners," and through intense emulation of their music, new techniques, improved instruments, and elaborated musical forms developed. Persian lute tuning was adopted for the lute (*ʿūd*), which became the classical instrument of the Arabs. Melodies and rhythms were regulated by a modal system that was later codified. Among the most famous female musicians was ʿAzza al-Maylāʾ, who excelled in *al-ghināʾ ar-raqīq,* or "gentle song." Her house was the most brilliant literary salon of Medina, and most of the famous musicians of the town came under her tutelage. Also famed were the female musician Jamīla, around whom clustered musicians, poets, and dignitaries; the male musician Ṭuways, who, attracted by the melodies sung by Persian slaves, imitated their style; and Ṣāʾib Khāthir, the son of a Persian slave. Songs were generally accompanied by the lute (*ʿūd*), the frame drum (*duff*), or the percussion stick (*qaḍīb*).

Sophisticated secular music

The Umayyad and ʿAbbāsid dynasties: classical Islāmic music. Under the Umayyad caliphate (661–750) the classical style of Islāmic music developed further. The capital was moved to Damascus (in modern Syria) and the courts were thronged with male and female musicians, who formed a class apart. Many prominent musicians were Arab by birth or acculturation, but the alien element continued to play a predominant role in Islāmic music. The first and the greatest musician of the Umayyad era was Ibn Misjaḥ, often honoured as the father of Islāmic music. Born in Mecca of a Persian family, he was a musical theorist and a skilled singer and lute player. Ibn Misjaḥ traveled to Syria and Persia, learning the theory and practice of Byzantine and Persian music and incorporating much of his acquired knowledge into the Arabian art song. Although he adopted new elements such as foreign musical modes, he rejected other musical traits as unsuitable to Arabian music. Knowledge of his contributions is contained in the most important source of information about music and musical life in the first three centuries of Islām. This is the 10th-century *Kitāb al-Aghānī,* or "Book of Songs," by Abū al-Faraj al-Iṣbahānī. In the 8th century Yūnus al-Kātib, author of the first Arabic book of musical theory, compiled the first collection of songs. Other notable musicians of the period were Ibn Muḥriz, of Persian ancestry; Ibn Surayj, son of a Persian slave and noted for his elegies and improvisations (*murtajal*); his pupil al-Gharīḍ, born of a Berber family; and the Negro Maʿbad. Like Ibn Surayj, Maʿbad cultivated a special personal style adopted by following generations of singers.

Further influence of conquered peoples

By the end of the Umayyad period, the disparate elements of conqueror and conquered were fused into the style of classical Islāmic music. With the establishment of the ʿAbbāsid caliphate in 750, Baghdad (in modern Iraq) became the leading musical centre. The ʿAbbāsid caliphate is the period of the Golden Age in Islāmic music. Music, obligatory for every learned man, was dealt with in varied aspects—among them virtuosity, aesthetic theory, ethical and therapeutic goals, mystical experience, and mathematical speculation. The artist was required to possess technical proficiency, creative power, and almost encyclopaedic knowledge. Among the finest artists of the period were Ibrāhīm al-Mawṣilī and his son Isḥāq. Members of a noble Persian family, they were chief court musicians and close companions of the caliphs Hārūn ar-Rashīd and al-Maʾmūn.

The Golden Age

Isḥāq, a singer, composer, and virtuoso lutenist, was the outstanding musician of his time. A man of wide culture, he is credited with authorship of nearly 40 works on music, which were subsequently lost. According to the "Book of Songs," he is the originator of the earliest Islāmic theory of melodic modes. Called *aṣbiʿ* ("fingers"), it structured the modes according to the frets of the lute and the fingers corresponding to them. Indications above each song in the "Book of Songs" show the mode, the type of third (major, minor, or neutral), and often the rhythmic mode. (The third is the interval encompassing three notes of the scale. It can vary considerably in exact size without losing its character. Western music uses the major and the minor third; much non-Western and folk music also uses a neutral third, between the major and minor in size.) The neutral third, introduced into Islāmic music about this time, increased the number of melodic modes from eight to 12 by making more intervals available from which to build melodies. At this time the number of rhythmic modes varied from six to eight, their actual structure and content differing from author to author.

Isḥāq and Ibrāhīm al-Mawṣilī actively participated in the contemporary controversy between modernism, a Persian romantic style tending toward exuberance of embellishments, and Arabian classicism, characterized by simplicity and artistic severity. The Mawṣilīs represented the older classical tradition; the proponents of modernism were Ibn Jāmiʿ and the celebrated singer Prince Ibrāhīm ibn al-Mahdī.

Theoretical writings

In the second half of the 8th century, the extensive Islāmic literature of music theory began to flourish. Greek treatises were translated into Arabic, and scholars, who were acquainted with the Greek writings, began to devote books or sections of books to the theory of music. In their works they expanded, changed, improved, or shed new light on Greek musical theory. The well-known philosopher al-Kindī, who was deeply immersed in Greek learning, wrote more than 13 musical treatises, including the earliest Arabic musical treatise that is known to have survived. He also dealt with the theory of ethos (*ta'thīr*) and with cosmological aspects of music. Members of the Ikhwān aṣ-Ṣafā, an important 10th-century brotherhood, dealt also with these two themes and advanced a theory of sound that went well beyond ancient Greek theories. Philosophers such as al-Fārābī, author of the monumental *Kitāb al-musīqī al-Kabīr* ("Grand Book on Music"), and Ibn Sīnā (known in Europe as Avicenna) dealt with such topics as the theory of sound, intervals, genres and systems, composition, rhythm, and instruments, as did others such as as-Sarakhsī, his contemporary Thābit ibn Qurrah, and Avicenna's pupil Ibn Zaylā. The last important theorist to emerge during the ʿAbbāsid period was Ṣafī ad-Dīn, who codified the elements of the modal practice as it was then known into a highly sophisticated system. His achievement became the chief model for subsequent generations. In the numerous treatises written between the 13th and 19th centuries, the system devised by Ṣafī ad-Dīn was split into multiple local traditions.

Islāmic music in Spain. Parallel to the flourishing of music at the eastern centres of Damascus and Baghdad, another important musical centre developed in Spain, first under the survivors of the Umayyad rulers and later under the Berber Almoravids (rulers of North Africa and Spain in the 11th and 12th centuries) and Almohads, who expanded into Spain after the fall of the Almoravids. In Spain, encounter with different cultures stimulated the development of the Andalusian, or Moorish, branch of Islāmic music. The most imposing figure in this development is Ziryāb (fl. 9th century), a pupil of Isḥāq al-Mawṣilī, who, because of the jealousy of his teacher, emigrated from Baghdad to Spain. A virtuoso singer and the leading musician at the court of Córdoba, Ziryāb introduced a fifth string to the lute, devised a number of new forms of composition, and developed a variety of new methods of teaching singing in his well-known school of music. Musical activity spread to large towns, and Seville became a leading centre of musical-instrument manufacture.

Poetic and musical forms

New poetic forms were developed, such as the *muwashshaḥ* and the *zajal*, that were freer in rhyme and metre than the classical *qaṣīdah* or formal ode. These innovations in prosody opened the way to further musical developments. Especially important was the *nawbah* ("suite"), a form that included songs and instrumental music, free or metrical, that were linked together by melodic mode and rhythmic patterns. The 24 traditional *nawbah*s were invested with symbolic and cosmological significance. After the expulsion of the Muslims from Spain in 1492 this musical tradition was transported to North African centres, where it partially survived.

After the Mongol invasion of Baghdad in 1258 and the Spanish reconquest of Granada in 1492, the magnificence of Islāmic culture gradually waned. Music continued to be cultivated, receiving new influences from Mongol and Turkmen conquerors. Persia enjoyed artistic independence for about 450 years, until 1918; but during this period a huge area, from the Balkans to Tunisia, was submitted to a strong Turkish influence, which itself was heavily influenced by Arab and Persian music.

The modern period. From the beginning of the 19th century, Islāmic music was affected by the intensification of contacts and relationships with Western music. For the first time Islāmic music existed in juxtaposition with Western music. For example, European composers and musicians were summoned to create military bands and conservatories in Turkey (1826) and in Persia (1856), and Giuseppe Verdi's opera *Aida* inaugurated the opera house in Cairo in 1871. Expanding contact with Western music caused certain alterations in traditional musical styles. There was a widespread musical renaissance, with two main centres: the leading school in Egypt was open to modernism and Western influences, while in Syria and Iraq traditional music was supported. Music in Syria and Iraq, together with North African, Iranian, and Turkish music, remained restricted to its own periphery. The Egyptian school developed Middle Eastern music in what can be called the mainstream style; and this music was widely diffused through the media of radio, television, recordings, and the cinema. Mainstream music borrowed instruments such as the cello, saxophone, and accordion; melodies and rhythms from European serious and light music; the concept of large ensembles; and the use of electronic amplification. Emphasis shifted from the display of individual virtuosity and personal creativity to performance as an ensemble, and the use of short songs underscored the separation, rather than the traditional union, of composer and performer. Classical and local genres coexist, however, with the innovative mainstream style.

Persian art music continues to be organized into 12 traditional modes, or *dastgāh*, each of which contains a repertory of from 20 to 50 small pieces called *gūsheh*s ("corners"). In performance of instrumental and vocal music, the artist improvises on the chosen *gūsheh*s of a *dastgāh* in a specific order.

Vocal music still predominates even in countries such as Iran, in which instrumental music is cultivated independently. Thus almost all of the Near Eastern musicians who are well known are singers; those particularly influential in the modern renaissance, in chronological order, include ʿAbduh al-Ḥamūlī, Dāhūd Ḥussnī, Sayyid Darwīsh, ʿAbd al-Wahhāb, Umm Kulthūm, Farid al-Aṭrash, Fayrouz, Rashid al-Hundarashi, Ṣadīqa al-Mulāya, and Muḥammad al-Gubanshi.

Theoretical developments

Modern Arab theorists also have produced valuable treatises. For example, the 19th-century theorists Michel Muchaqa of Damascus and Mohammed Chehab ad-Dīn of Cairo introduced the theoretical division of the scale into 24 quarter tones. In 1932 the international Congress of Arabian Music was held in Cairo, providing a forum for current analysis of subjects such as musical scales, modes, rhythms, and musical forms. (A.Sh.)

Dance and theatre

The performing arts have received comparatively little attention in the otherwise rich literature of the Islāmic peoples. This is most probably a result of the suspicions entertained by some orthodox Muslim scholars concerning the propriety of the dance and the theatre. Because this applies particularly in relation to the vexing theological question of human portrayal and its connection with idolatry, the performing arts have traditionally been regarded by the faithful with more than usual caution. Even as late as the 19th and early 20th centuries, most research on the subject, in what may loosely be called the Islāmic world, was carried out by Western scholars, chiefly from European nations; and only in the 20th century have indigenous scholars started publishing significant research on the subject.

There are no known references to the dance or theatre in pre-Islāmic Arabia, although nomad tribes were probably acquainted with the dance. The Islāmic peoples themselves seem to have developed this particular art form less than they did music or architecture; and, in addition to medieval Islām's cool attitude toward dance and theatre as art forms, it must be added that most women, leading a life of seclusion, could hardly be expected to play an active part in them. Nevertheless, there has been an active tradition of folk dance in most Islāmic countries, in addition to dancing as an entertainment spectacle and, particularly in Persia, as an art form. A ritual dance was instituted in

Dance as entertainment for the aristocracy, shown in "A Festive Party," manuscript illumination from the *Masnavī* of Jalāl ad-Dīn ar-Rūmī, AD 1295–96. In the British Museum (MS. OR. 7693, fol. 225 b.).

By courtesy of the trustees of the British Museum; photograph, J.R. Freeman & Co. Ltd.

the Ṣūfī mystical order of the Mawlawīyah (Mevleviyah) in Turkey. The dance, performed by dervishes (members of the mystical order), is considered to be a manifestation of mystical ecstasy rather than an entertainment or an expression of aesthetic urges.

The theatre has not flourished as a major art under Islām, although as a form of popular entertainment, particularly in mime and shadow-puppet shows, it has persisted vigorously. Nevertheless, the theatre with live actors received support from the Ottomans in Turkey, and a live popular drama has been strong in Persia, where a passion play also took root. Otherwise, the theatrical record of Islām is meagre. Moreover, few neighbouring peoples had a well-developed theatre of their own; hence, outside stimulus was lacking, and the Islāmic disapproval of idolatry was so intense that, when the shadow theatre evolved in the East in the late Middle Ages, the puppets were regularly punched with holes to show that they were lifeless. Nonetheless, drama has had some ties with religion, as in Iran and other areas where the Shīʿite branch of Islām is concentrated. Here a passion play developed, rooted in traumatic memories of the bloody warfare of Islām's early years. This was a local phenomenon, uninfluenced by Christian Europe, and, though stereotyped, it movingly reenacted Shīʿite martyrdom.

Popular theatre

A popular theatre, frequently including dance, evolved independently from about the 17th century in some Muslim countries. West European and, later, U.S. influences were largely the main factors in the development of an artistic theatre in the 19th and 20th centuries. But conservative Muslims have consistently disapproved of theatre, and in Saudi Arabia, for example, no native theatrical establishment exists. In such an atmosphere, women's parts were at first taken by men; later, Christian and Jewish women took the roles, and only in the 20th century have Muslim women participated.

TYPES AND SOCIAL FUNCTIONS OF DANCE AND THEATRE

The dance. Folk dancing existed among medieval Islāmic peoples; but such sources as exist are mainly concerned with artistic dance, which was performed chiefly at the caliph's palace by skilled women. The aristocracy was quick to imitate this patronage by providing similar performances, its members vying with one another on festive occasions. One of these dances, the *kurrağ* (some times called *kurra*), developed into a song and dance festival held at the caliph's court. Since the latter part of the 19th century, the dancing profession has lost ground to the performance of U.S., Latin-American, and western European dances in cabarets. In a reaction that set in after World War II, fervent nationalists have tried to create native dance troupes, revive traditional motifs in costume and interpretation, and adapt tribal figures to modern settings. Few traditional dances have survived unchanged; among those that have are the dervish dances, performed mainly in Turkey.

Folk dance. Though now performed and fostered chiefly as an expression of national culture, folk dances were long regarded as pure entertainment and were either combined with theatrical shows or presented alone. Dance performances, accompanied by music, took place in a special hall or outdoors; many dancers, particularly the males, were also mimes. Sometimes the dance enacted a pantomime, as in Turkey, of physical love or of a stag hunt, representing the pursuit of a suspicious husband deceived by his wife.

Folk dance, except in Iran, has almost always been mimetic or narrative, a tradition still fostered by many tribes.

Dance as entertainment. The Turks considered dancing as a profession for the low-born; as a result, most dancers

By courtesy of the University of Chicago Library

Dervishes dancing in a *tekke*, engraving by J. Fougeron from *A Tour to the East in the Years 1763 and 1764, with Remarks on the City of Constantinople and the Turks*, by Frederick Calvert, 6th Baron Baltimore, 1767.

were members of minority groups: mostly Greeks, Jews, and Armenians. This judgment has usually applied to the status of professional dancers and indeed to most professional entertainers at most periods and in most societies until modern times. In 19th-century Egypt, both male and female dancers were regarded as public entertainers. Many of the women entertainers (*ghawāzī*) belonged to a single tribe and were usually considered little better than prostitutes. The erotic element in dancing has become focused in the belly dance, which has become the leading form of exhibition dance in modern Turkey and the Arab countries.

The mimetic tradition of folk dance has blended well in countries of the Sunnite persuasion with comedy and with the passion-play tragedy in Shīʿite countries. In recent years, however, the theatre has been increasingly divorced from the dance, with most plays being consciously modeled on European patterns; only in the operetta does the old combination remain.

Dance as an art form. In pre-Islāmic times in Iran, dance was both an art form and a popular entertainment. There are pictures of dancers in miniatures, on pottery, and on walls, friezes, and coins. Some of these ancient dances lived on partially in tribal dances, but again, under Islām's restrictions on women, the art became a male monopoly. Women were permitted to dance in private, however, as in the harem. Iran is perhaps the only Muslim country with a tradition of dance regarded as an art form. When revived after World War II, folk dancing was encouraged and adapted for the foundation of a national ballet. Muslim orthodoxy's very uncertainty over the exact status of the artistic dance ensured that it was always considered as an adjunct to music. Accordingly, although there are many detailed treatises on Islāmic music, none is available on dance.

Dervish dancing. There is one outstanding example of pure dance: that of the whirling dervishes, an art that has been practiced for more than seven centuries. The procedure is part of a Muslim ceremony called the *dhikr,* the purpose of which is to glorify God and seek spiritual perfection. Not all dervish orders dance; some simply stand on one foot and move the other foot to music. Those who dance or, rather, whirl are the Mawlawī dervishes, an order that was founded by the Persian poet and mystic Jalāl ad-Dīn ar-Rūmī at Konya, in Anatolia, in the 13th century.

The performance, for which all of the participants don tall, brown, conical hats and black mantles, takes place in a large hall in the *tekke,* the building in which the dervishes live. The dervishes sit in a circle listening to music. Then, rising slowly, they move to greet the *shaykh,* or master, and cast off the black coat to emerge in white shirts and waistcoats. They keep their individual places with respect to one another and begin to revolve rhythmically. They throw back their heads and raise the palms of their right hands, keeping their left hands down, a symbol of giving and taking. The rhythm accelerates, and they whirl faster and faster. In this way they enter a trance in an attempt to lose their personal identities and to attain union with the Almighty. Later they may sit, pray, and begin all over again. The *dhikr* ceremony always ends with a prayer and a procession.

Reed Consumer Books Picture Library

Women dancing, from a reconstruction of a wall painting from the harem of the Jausaq al-Khāqānī Palace, Sāmarrā', Iraq, 833–841. In the Museum of Turkish and Islamic Art, Istanbul.

The theatre. In lands where the Sunnite sect was strong, mime shows were frequent and popular attractions during the later Middle Ages. The Ottoman sultans were accompanied on military campaigns by their own troupe of actors; and, as the Ottoman Empire grew larger and richer, the court became ever more partial to entertainment, whether at the accession of a sultan, a royal wedding, a circumcision, an official visit, or a victory. On such occasions, dances and theatrical performances played their part along with parades, fireworks, music, mock fights, and circus performances in one huge, sumptuous pageant. This lavishing of entertainment reached a height of splendour that the admiring Ottoman aristocracy strove to imitate throughout the empire. In Arabia and North Africa, popular shows on a lesser scale were performed in the open air. Another aspect of the Islāmic theatre was represented in the shadow plays, which were given chiefly to pass the time during the month of fasting, Ramaḍān (the sacred ninth month of the Muslim year).

Actors at the Ottoman court

Among Shīʿites the passion play was regularly performed, by both professional and amateur actors. The performance always took place during the first 10 days of the month of Muḥarram (the first in the Muslim year), the period when the suffering and death of the descendants and relatives of the fourth caliph ʿAlī were commemorated. For generations this largely theatrical event served as a focal point of the year, gripping audiences in total involvement with its blend of symbolism and realism.

Mime shows. In the medieval Muslim theatre, mime shows aimed to entertain rather than to uplift their audiences. Regrettably, few mime shows were recorded in writing, and those that were recorded were set down primarily to serve as guidelines for directors, who might tamper with the wording, as in the improvisation of the Italian commedia dell'arte. Some plays were on historical themes, but preference was for comedies or farces with an erotic flavour. The audience was largely composed of the poor and uneducated.

A rudimentary theatrical form, the mime show long enjoyed widespread popularity in Anatolia and other parts of the Ottoman Empire. Called *meddah* (eulogist) or *mukallit* (imitator) in Turkish, the mimic had many similarities to his classical Greek forerunners. Basically he was a storyteller who used mimicry as a comic element, designed to appeal to his largely uneducated audience. By gesture and word he would imitate animals, birds, or local dialects; he was very popular in Arabic- and Turkish-speaking areas. Even today, he has not been wholly supplanted in the Islāmic world by literacy or by such modern entertainments as radio, television, and the cinema. Sometimes several *meddah*s performed together, and this may have been the source of a rural theatrical performance.

Ortaoyunu. The *ortaoyunu* (middle show) was the first type of genuine theatre the Turks, and possibly other Muslim peoples, ever had. The Ottoman sultans provided subsidies for *ortaoyunu* companies of actors, who consequently became generally accepted; also some were retained by the princes of the Romanian principalities under Ottoman rule. The fact that they continued to enjoy popularity to World War I may be explained by their simple dramatic appeal, which was coupled with sharp satire of the well-to-do and the ruling classes (but hardly ever of Islām). This irreverence frequently resulted in fines and imprisonment for the actors, but it never produced a basic change of style.

The first real theatre

During the 19th and 20th centuries, the *ortaoyunu* was generally performed in an open square or a large coffeehouse. There was no stage, and props were simple: they generally comprised a table or movable screen, while other

Ortaoyunu theatre, painting by Muazzez. In the collection of Dr. Metin And.
Metin And

objects were represented by paintings glued on paper. An orchestra of about four musicians enlivened the show and gave the performers, who were all male, their cues. Roles were generally stereotyped, with stock characters, such as a dandy, the foreign physician, and regional types (Kurds, Albanians, Armenians, Arabs, and Jews) quarreling and fighting in slapstick style. Mimicry was important, and some actors changed roles and costumes. The plot was flimsy, a mere frame for the dialogue, which was itself frequently improvised.

The marionette theatre. In comparison with *ortaoyunu,* the marionette theatre, although popular in Turkistan (under the name of *çadir hayâl*) and other parts of Muslim Central Asia, never really caught on in the Ottoman Empire.

Shadow plays (Karagöz). On the other hand, the shadow play had been widely popular for many centuries in Turkish- or Arabic-speaking countries. Its essence, like that of the mime shows, was entertainment without moral import; and few plays were recorded in writing beyond a sketch of the action. Most were comedies and farces that were performed for the enjoyment of an audience that was, for the most part, very poor and uneducated.

In Turkey, the Karagöz (a character, "Black-eye") theatre was the prevalent form of shadow play. This art apparently came from China or perhaps from Southeast Asia, as the French term *ombres chinoises* indeed hints, though the prevailing element of the grotesque was probably inherited from ancient Greece by way of Byzantium. The Karagöz was well known in Turkey during the 16th century but was so fully developed that it must have been introduced much earlier, and it quickly spread from Syria to North Africa and the Greek islands. Its performers were in great demand at the sultan's court as well as elsewhere, and they soon organized their own guild. Since only the framework of the play was sketched in writing, there was scope for a great deal of impromptu wit, and Karagöz shows, like the *ortaoyunu,* were inevitably satirical. But with the coming of motion pictures the Karagöz declined, and performances are now mostly confined to the month of Ramaḍān.

In the traditional performance of the Karagöz, the stage is separated from the audience by a frame holding a sheet; the latter has shrunk over the years from about six by 7½ feet (1.8 by 2.3 metres) to about three by two feet (0.9 by 0.6 metres). The puppets, which are flat and made of leather, are controlled by the puppeteers with rods and are placed behind the screen. An oil lamp is then placed still farther back so that it will throw the puppets' shadows onto the screen.

Traditional characters

A standard shadow play has three main elements: introduction, dialogue, and plot. The introduction is fairly stereotyped and consists of an argument and usually a quarrel between Karagöz and Hacivat, the two most com-

Marc Riboud—Magnum

Karagöz shadow puppets.
From left: Yahudi (the Jew) with donkey, Karagöz, Zenne (the woman), and Tasuz Deli Bekir.

mon characters. The former is a simple, commonsense fellow, while the latter is more formal and polished, if shallow and pedantic. The dialogue between the two varies with the occasion but always contains impromptu repartee, though most puppet masters have at least 28 different plots in stock—a different one for each night of Ramaḍān. Some are historical, many ribald, but all are popular entertainment. Additional characters or animals may be introduced, calling for great skill on the part of the puppet master and his assistant in manipulating several simultaneously, as well as in reciting the text in changing tones and playing music. Some have one or two musicians to help.

Mimicry and caricature, while essential to both the *meddah* and the *ortaoyunu,* are technically more developed in the shadow play. Here entire productions are based on a comedy of manners or of character. In addition to the stock characters from various ethnic groups, there is, for example, the drug addict who wraps his narcotic in dissolving gum before the fast begins so as not to sin, the light-headed Turk ("he who eats his inheritance") who is a prodigal and a debauchee, the highway robber, the stutterer, and the policeman.

Karagöz is the most frequently performed but not the sole type of shadow play in Muslim countries. In Egypt a shadow theatre is known to have existed as early as the 13th century, long before records of Karagöz shows were kept in Turkey. A physician, Muḥammad ibn Dāniyāl, wrote three shadow plays that have survived. They were performed in the 13th century and display humour and satire and the lampooning of match-making and marriage. These plays also introduce a parade of popular contemporary characters, many of whom earn their living in shady or amusing trades. A positively phallic element is as evident here as it is in the Karagöz.

Iranian popular theatre. Popular theatre existed among the Iranians, who were proud of a long-lived cultural tradition and preserved their national language under Arab domination: indeed, even their branch of Islām, Shīʿism, set them apart from the Sunnism of the majority of Islām. The Ottomans' failure to conquer Iran increased competition between the respective intellectual elites. Iran had inherited a considerable theatrical tradition from pre-Islāmic times; it is not surprising that a popular comic theatre flourished there. The central figure of this theatre was the *Katchal Pahlavān* (or "bald actor"), and mimicry was important, both in comedy and in pantomime. The *Baqqal-Bāzī* ("Play of the Grocer"), in which a grocer repeatedly quarrels with his good-for-nothing servant, is a typical example of the popular comic tradition. The marionette theatre, or *Lobet-Bāzī*, while using Iranian puppets, was similar to its Turkish counterpart. At least five puppets appeared, and singing was an integral part of a production that sometimes resembled Italian and French puppet shows. The *ortaoyunu,* particularly in the region of Azerbaijan, is almost identical with the Turkish of the same name. The shadow play in Iran, however, has always been less popular and obscene than the Ottoman or Arab Karagöz.

The comic theatre in Iran

Passion plays (taʿziyah). Quite different was the passion play, derived mainly from early Islāmic lore and assembled as a sequence of tragedies representing Shīʿite martyrdom. Both shadow and passion play were interlarded with musical prologues, accompaniment, and interludes; but these were not necessarily an integral part, serving rather to create a mood.

A preoccupation with religion is characteristic of Persian theatrical performances, and, during the first 10 days of the month of Muḥarram, the martyrdom of ʿAlī's descendants at the hands of the Umayyads is reenacted. Although these shows are also performed among Shīʿite Turks in Central Asia and Shīʿite Arab communities in Iraq and elsewhere, Iran is their centre. Some plays are satirical, directed against wrongdoers, but most form a set of tragedies, performed as passion plays on these 10 successive days. Named *taʿziyah* ("consolation"), this type of drama is an expression of Persian patriotism and, above all, of piety, both elements combining in an expression of the national religion, Shīʿism.

Background of the *taʿziyah*

In order to understand the mood of the *taʿziyah* it is necessary to remember that storytellers in Iran recite the gruesome details of the martyrdom of Ḥasan, Ḥusayn, and other descendants of ʿAlī all year long. Thus prepared, people swell the street processions during the days of Muḥarram, chain themselves, flagellate their bodies, and pierce their limbs with needles, shouting in unison and carrying images of the martyrs, made of straw and covered with blood—contrary to the injunctions of Islām. Sometimes men walk in the processions with heads hidden and collars bloodied, all part of a pageant dating from the 9th or 10th century. Its peak is reached daily in the play describing the martyrdom of ʿAlī's family and entourage, which used to be presented in the large mosques, but which, when the mosques proved too small, was given a special place. The roles of reciter of the martyrdom and of participant in a procession have blended over the years to produce the *taʿziyah* play, in which the reciters march in procession to the appointed place and there recite their pieces, which can be considered as a prologue before the play itself begins.

The chief incidents narrated in the *taʿziyah* are not necessarily presented in chronological order, but in any case the *taʿziyah* texts (manuscripts from the 17th and 18th centuries, thenceforth, printed texts) give an inadequate impression of their forceful effect. Indeed, the audience identifies itself so closely with the play that foreigners have, on occasion, been manhandled. Since half of the actors play the supporters of the ʿAlids and half play their opponents, the latter are sometimes attacked and beaten at the end of the play. The decor, too, is half-realistic and half-symbolic: blood is real, yet sand is represented by straw. The stage effects are frequently overdone, and this clearly further excites the audience. For instance, Ḥusayn's gory head is made to recite holy verses; or an armless warrior is seen to kill his opponent with a sword he holds in his teeth. The horses are real, although most of the other animals are played by humans. In general, the actors, though chiefly nonprofessional, infect the audience with their enthusiasm and absorption.

DANCE AND THEATRE IN MODERN TIMES

Developments in dance. Insofar as dance is related to the modern theatre, there is little difference between Muslim production and its European or American counterpart. Dance and drama are combined according to the artistic needs of the production or the personal tastes of the producer and director. Perhaps more important is the dance itself, independently performed as artistic self-expression. The geographical centre of folk dance is in the area east of the Mediterranean, though remnants of other cultures have survived. There are Balkan traces in western and northern Turkey, for example, and Berber and even black African traces in Morocco and elsewhere in North Africa.

Arab countries. In some Arab countries, dancing is popular, varying in town, village, or with nomad tribe. In the town, dancing is generally reserved for special occasions, chiefly Western social dances. On the other hand, villages have such favourites as the *dabkah.* The *dabkah* is danced mainly by men and is quite common in festivities in the area between northern Syria and southern Israel; for instance, the Druzes (sectarian Arab communities located in Lebanon, Syria, and Israel) are very fond of it. The performers dance in a straight line, holding handkerchiefs high in the air, while the first dancer in the row gives the sign for stepping or jumping. Among the Bedouin almost any pretext suffices for dancing, although since the mid-20th century dancing has been practiced most often at weddings and similar festivities. Usually two male dancers, or two rows of male dancers, repeatedly advance toward each other or the audience and retire. To this basic figure, there are numerous variations that give the different dances their names.

Turkey. The Turks are also lovers of music and dance and when they meet frequently sing and dance. There is no single national dance popular throughout the country; dances vary in the numbers required, some being for solo performance, others designed for pairs or groups, though nearly all have instrumental accompaniment. As illustration of the possibilities of a basic step, there are at least 40 variations of the group dance known as *bar*, a chain dance. Again, several folk dances have characteristics akin to pantomime, breaking up into five main types of imitation: village life, nature, combat, courtship, and animals or birds.

Opera is popular in Turkey, reflected in a long tradition of invitations to foreign companies, and the musical theatre, which frequently includes dancing, is also widespread. On the other hand, classical ballet was unknown until a school of ballet was opened by foreign teachers with government encouragement. Although most of the ballet performances are in Istanbul, they are well received on tour.

Iran. In Iran a national dance company was formed with government support after World War II, and ancient customs were revived. Until it was closed in 1979, the Iranian ballet company was outstanding in the Muslim world, drawing on ancient war dances, fire-priest dances, dervish dances, and tribal folklore, as well as on scenes and decor from painting, sculpture, and the rich imagery of classical Persian poetry. Various folk dances are likewise performed all over Iran; they are accompanied by music and reflect local traditions and customs. Some are mimetic, others erotic, others, again, war dances (chiefly in the mountain areas) and comic dances (usually with masks). Many of these are dying out as new tastes and customs evolve, and Iranian dance companies have tried to preserve some of these dying forms.

Iranian ballet

The contemporary theatre. The modern Muslim theatre is almost wholly a western European importation, unconnected with the traditional medieval theatre, which has almost completely disappeared, although there are vestiges of it.

Arab countries. Contemporary Arabic theatre owes much to the imaginative daring of the Naqqāsh family in 19th-century Beirut, which was then under Turkish rule. Significantly, they were Christians, then better educated and more cosmopolitan than Muslims, and they had the advantages of Beirut's contacts with Europe and position as the headquarters of missionary activity. A Beirut Maronite (a Roman Catholic following the Syrio-Antiochene rite, widespread in the area), Mārūn an-Naqqāsh (died 1855), who knew French and Italian as well as Arabic and Turkish, adapted Molière's *L'Avare* ("The Miser") and presented it on a makeshift stage in Beirut in 1848. He did so before a select audience of foreign dignitaries and local notables, and he wrote his play in colloquial Arabic and revised the plot to suit the taste and views of his audience. Further, he changed the locale to an Arab town and Arabicized the names of the participants. Other touches included instrumental and vocal music and the playing of women's roles by men, in the traditional manner. The above features characterized the Arabic theatre for about half a century. An-Naqqāsh, together with his family, composed and presented two other musical plays, one based on Molière's *Tartuffe,* the other on the story, in *The Thousand and One Nights,* of Abū al-Ḥasan, who became caliph for a day.

Soon the main centre of Arabic theatre moved to Egypt, whose comparatively tolerant autonomy offered an atmosphere for literary and artistic creativity more congenial than other parts of the Ottoman Empire. Syrian and Lebanese intellectuals and actors emigrated there, particularly after the anti-Christian riots of 1860 in Syria. Though a somewhat crippled Arabic theatre continued in Syria, its influence was carried into Egypt by émigrés and later spread to other Arabic-speaking regions. The number of theatres, a potentially large public, the munificence of Egypt's rulers, increasing prosperity under British rule after 1882, and increasing education soon made Egypt the centre of Arabic theatre, a position it has successfully maintained since.

Egyptian theatre

The colloquial Arabic of Egypt was increasingly employed in the theatre, and several companies toured the country and neighbouring parts. The composition of these companies was fluid, for the actors were prone to be fickle in their loyalties; nevertheless, certain types of Egyptian theatre can be discerned in the late 19th century and during the early 20th. Some, like the company of Salāmah Ḥijāzī, used music to such an extent that their productions approached being labelled opera or operetta. Others, like that of ʿAlī al-Kassār, specialized in downright farce, expressed in revue form, with a Nubian hero, the "Barbarin," who made a specialty of ridicule and mimicry. Yet others, like the company of Najīb ar-Rīḥānī, oscillating between outright farce and comedy, skillfully depicted contemporary Egyptian manners; in particular, Najīb ar-Rīḥānī created a character called Kish-Kish Bey, whose misadventures and unsolicited advice on every subject have made him a classic creation. A conventional theatre sprang up in Egypt, too, catering to a growing number of intellectuals and presenting dramas and tragedies in polished, literary Arabic. Its chief exponent was Jūrj Abyaḍ, who had spent time studying acting in Paris. In contrast, Yūsuf Wahbī's National Troupe performed realistic plays, usually dramas or melodramas, using either colloquial or literary Arabic and sometimes a combination of both.

The plays performed by the Egyptian troupes and others in Arabic-speaking lands developed through three overlapping but distinguishable stages: adaptations, translations, and original plays. Adaptations came first in the 19th century (see above). Translations of established works appealed to a discriminating public, but original plays, part of the evolution of modern Arabic literature, reflected a growing interest in political and social problems. The decline of foreign influence and the arrival of political independence encouraged creativity, which, however much under European influence, has some original works to its credit. Two 20th-century Arabic playwrights, both Egyptian, are Tawfīq al-Ḥakīm, a sensitive shaper of both social and symbolic dramas, and Maḥmūd Taymūr, a novelist and comedy writer who strikes deep into Egypt's social problems.

Turkey. The development of the modern Turkish theatre strongly resembles its Arabic counterpart. In Istanbul, theatrical performances were not unusual among the diplomatic and international set, and some local Turks were acquainted with them. Nonetheless, Turkish plays for live actors—barring *ortaoyunu*—date only from 1839. The first Turkish playhouse was built in Pera (now Beyoğlu), significantly in the middle of the foreign and embassy quarter of Istanbul. Many of the actors were members of non-Muslim minorities, such as the Armenian; and the first plays presented in Turkish were adaptations from the French, chiefly Molière. They were done during the 1840s, when music was an important item.

The Gedik Paşa Theatre, which is named for the area in Istanbul where it was located, was the first theatre in which Turkish plays were produced by native actors speaking in Turkish. The actors received a salary, and local writers presented their own plays. Originally built for foreign companies, the theatre was reconstructed in 1867 and reopened in 1868 for a Turkish company headed by an Armenian, Agop, who was later converted to Islām and changed his name to Yakup. For almost 20 years the Gedik Paşa Theatre was the dramatic centre of the city; and plays in translation were soon followed by original plays, several with a nationalist appeal, such as Namık Kemal's *Vatan yahut Silistre* (*Fatherland*), which was first produced in 1873. The actors had to struggle against prejudice and the playwrights against censorship (some of them were imprisoned or exiled), but the Turkish theatre spread beyond Istanbul in the 1870s and 1880s to such places as Adana (in southern Anatolia) and Bursa (just south of Istanbul, across the Sea of Marmara).

The Gedik Paşa Theatre

After the Young Turk Revolution of 1908, censorship was not relaxed, but interest in the theatre grew, particularly over political matters; and plays about the new constitution were written and performed. After the foundation of the Turkish Republic in 1923, the state subsidized several theatre companies and a school for dramatic arts, and an opera house was built in Ankara. Official support not only gave financial encouragement but also implied a change of attitude over such matters as the participation of Muslim women in productions.

By the middle of the 20th century, theatrical life was mostly centred on Istanbul and Ankara, although theatres and companies continued in the small towns too.

A growing number of original plays, some of which were influenced by American literature, have been written and produced; the standard has been higher than it was before World War I, when Turkish poetry and fiction were rather more impressive than the drama. Subjects, too, have been more diverse since that time. To topics such as the position of women, marriage and divorce, and the character of Islāmic institutions—all popular under the Ottomans—have been added the Greco-Turkish War, education, village conditions, secularization, class struggle, and psychological problems. The Dormen Theatre was founded in Istanbul in 1955 by Haldun Dormen; in the 1971 World Theatre season in London the company performed *A Tale of Istanbul,* a comedy that included elements of folklore, a puppet show, singing, and a belly dance. The Dormen Theatre also produces 20th-century Western plays.

Western European and Russian influences in Iran

Iran. In Iran the birth of the modern theatre dates from the second half of the 19th century. Adaptations and translations from European plays appeared in Persian, often with the location and names suited to Iran. Molière, again, was a favourite and western European influence considerable, though Russian literature also left its mark, particularly in Azerbaijan, whose northern population had a chance to watch Russian actors during World War I.

Playwrights began to write original plays almost at once; one of the earliest playwrights was an Azerbaijani, named Akhundof, living in the Caucasus. He wrote seven comedies ridiculing Persian and Causasian Muslim society; all were translated into Persian and printed in 1874. Other plays likewise showed pronounced yearnings for social reform presented in a satirical style; some of these were published in a magazine called *Tyatr* ("Theatre"), which first appeared in 1908. Another type was the patriotic play, extolling Iran's history.

Some pre-World War I pieces were designed for reading rather than production. They were performed usually in schools, but there were hardly any professional actors, and the stage and props were very simple. After World War I, suitable halls were built in Tehrān and other cities, but the iron hand of Reza Shah (1925–41) curtailed development through continuous censorship and surveillance. After 1942 many new companies were formed, and there was speedy development, with growing interest in social and political subjects, though competition from foreign films was considerable. The revolutionary Islāmic regime established in 1979 severely curtailed theatrical activity.

(J.M.L.)

Visual arts

In order to answer whether or not there is an aesthetic, iconographic, or stylistic unity to the visually perceptible arts of Islāmic peoples, it is first essential to realize that no ethnic or geographical entity was Muslim from the beginning. There is no Islāmic art, therefore, in the way there is a Chinese art or a French art. Nor is it simply a period art, like Gothic art or Baroque art, for once a land or an ethnic entity became Muslim it remained Muslim, a small number of exceptions like Spain or Sicily notwithstanding. Political and social events transformed a number of lands with a variety of earlier histories into Muslim lands. But, since early Islām as such did not possess or propagate an art of its own, each area could continue, in fact often did continue, whatever modes of creativity it had acquired. It may then not be appropriate at all to talk about the visual arts of Islāmic peoples, and one should instead consider separately each of the areas that became Muslim: Spain, North Africa, Egypt, Syria, Mesopotamia, Iran, Anatolia, India. Such, in fact, has been the direction taken by some recent scholarship. Even though tainted at times with parochial nationalism, the approach has been useful in that it has focused attention on a number of permanent features in different regions of Islāmic lands that are older than and independent from the faith itself and from the political entity created by it. Iranian art, in particular, exhibits a number of features (certain themes such as the representation of birds or an epic tradition in painting) that owe little to its Islāmic character since the 7th century. Ottoman art shares a Mediterranean tradition of architectural conception with Italy rather than with the rest of the Muslim world.

Such examples can easily be multiplied, but it is probably wrong to overdo their importance. For if one looks at the art of Islāmic lands from a different perspective, a totally different picture emerges. The perspective is that of the lands that surround the Muslim world or of the times that preceded its formation. For even if there are ambiguous examples, most observers can recognize a flavour, a mood in Islāmic visual arts that is distinguishable from what is known in East Asia (China, Korea, and Japan) or in the Christian West. This mood or flavour has been called decorative, for it seems at first glance to emphasize an immense complexity of surface effects without apparent meanings attached to the visible motifs. But it has other characteristics as well; it is often colourful, both in architecture and in objects; it avoids representations of living things; it gives much prominence to the work of artisans and counts among its masterpieces not merely works of architecture or of painting but also the creations of weavers, potters, and metalworkers. The problem is whether these uniquenesses of Islāmic art, when compared to other artistic traditions, are the result of the nature of Islām or of some other factor or series of factors.

These preliminary remarks suggest at the very outset the main epistemological peculiarity of Islāmic art: it consists of a large number of quite disparate traditions that, when seen all together, appear distinguishable from what surrounded them and from what preceded them through a series of stylistic and thematic characteristics. The key question is how this was possible, but no answer can be given before the tradition itself has been properly defined.

Such a definition can only be provided in history, through an examination of the formation and development of the arts through the centuries. For a static sudden phenomenon is not being dealt with, but rather a slow building up of a visual language of forms with many dialects and with many changes. Whether or not these complexities of growth and development subsumed a common structure is the challenging question facing the historian of this artistic tradition. What makes the question particularly difficult to answer is that the study of Islāmic art is still so new. Many monuments are unpublished or at least insufficiently known, and only a handful of scientific excavations have investigated the physical setting of the culture and of its art. Much, therefore, remains tentative in the knowledge and appreciation of works of Islāmic art, and what follows is primarily an outline of what is known with a number of suggestions for further work into insufficiently investigated areas.

Media and techniques

Each artistic tradition has tended to develop its own favourite mediums and techniques. Some, of course, such as architecture, are automatic needs of every culture; and, for reasons to be developed later, it is in the medium of architecture that some of the most characteristically Islāmic works of art are found. Other techniques, on the other hand, acquire varying forms and emphases. Sculpture in the round hardly existed as a major art form, and, although such was also the case of all Mediterranean arts at the time of Islām's growth, one does not encounter the astounding rebirth of sculpture that occurred in the West. Wall painting existed but has generally been poorly preserved; the great Islāmic art of painting was limited to the illustration of books. The unique feature of Islāmic techniques is the astounding development taken by the so-called decorative arts—*e.g.,* woodwork, glass, ceramics, metalwork, textiles. New techniques were invented and spread throughout the Muslim world—at times even beyond its frontiers. In dealing with Islām, therefore, it is quite incorrect to think of these techniques as the "minor" arts. For the amount and intensity of creative energies spent on the decorative arts transformed them into major artistic forms, and their significance in defining a profile of the aesthetic and visual language of Islāmic peoples is far greater than in the instances of many other cultures. Furthermore, since, for a variety of reasons to be discussed later, the Muslim world did not develop until quite late the notion of "noble" arts, the decorative arts have reflected far better the needs and ambitions of the culture as a whole. The kind of conclusion that can be reached about Islāmic civilization through its visual arts thus extends far deeper than is usual in the study of an artistic tradition, and it requires a combination of archaeological, art-historical, and textual information.

Definition of a culture through an art

An example may suffice to demonstrate the point. Among all the techniques of Islāmic visual arts, the most important one was the art of textiles. Textiles, of course, were used for daily wear at all social levels and for all occasions. But clothes were also the main indicators of rank, and they were given as rewards or as souvenirs by princes, high and low. They were a major status symbol, and their manufacture and distribution were carefully controlled through a complicated institution known as the *ṭirāz.* Major events were at times celebrated by being depicted on silks. Many texts have been identified that describe the hundreds of different kinds of textiles that existed. Since textiles could easily be moved, they became a vehicle for the transmission of artistic themes within the Muslim world and beyond its frontiers. In the case of this one technique, therefore, one is not dealing simply with a medium of the decorative arts but with a key medium in the definition of a given time's taste, of its practical functions, and of the ways in which its ideas were distributed. The more unfortunate point is that the thousands of fragments that have remained have not yet been studied in a sufficiently systematic way, and in only a handful of instances has it been possible to relate individual fragments to known texts. When more work has been completed, however, a study of this one medium should contribute significantly to the commercial, social, and aesthetic history of Islām, as well as explain much of the impact that Islāmic art had beyond the frontiers of the Muslim world.

The following survey of Islāmic visual arts, therefore, will be primarily a historical one, for it is in development through time that the main achievements of Islāmic art can best be understood. At the same time, other features peculiar to this tradition will be kept in mind: the varying importance of different lands, each of which had identifiable artistic features of its own, and the uniqueness of some techniques of artistic creativity over others.

ORIGINS

Assimilation of earlier artistic traditions

Islāmic visual arts were created by the confluence of two entirely separate kinds of phenomena: a number of earlier artistic traditions and a new faith. The arts inherited by Islām were of extraordinary technical virtuosity and stylistic or iconographic variety. All the developments of arcuated and vaulted architecture that had taken place in Iran and in the Roman Empire were available in their countless local variants. Stone, baked brick, mud brick, and wood existed as mediums of construction, and all the complicated engineering systems developed particularly in the Roman Empire were still utilized from Spain to the Euphrates. All the major techniques of decoration were still used, except for monumental sculpture. In secular and in religious art, a more or less formally accepted equivalence between representation and represented subject had been established. Technically, therefore, as well as ideologically, the Muslim world took over an extremely sophisticated system of visual forms; and, since the Muslim conquest was accompanied by a minimum of destruction, all the monuments, and especially the attitudes attached to them, were passed on to the new culture.

The second point about the pre-Islāmic traditions is the almost total absence of anything from Arabia itself. While archaeological work in the peninsula may modify this conclusion in part, it does seem that Islāmic art formed itself entirely in some sort of relationship to non-Arab traditions. Even the rather sophisticated art created in earlier times by the Palmyrenes or by the Nabataeans had almost no impact on Islāmic art, and the primitively conceived ḥaram in Mecca, the only pre-Islāmic sanctuary maintained by the new faith, remained as a unique monument that was almost never copied or imitated despite its immense religious significance. The pre-Islāmic sources of Islāmic art are thus entirely extraneous to the milieu in which the new faith was created. In this respect the visual arts differ considerably from most other aspects of Islāmic culture.

This is not to say that there was no impact of the new faith on the arts, but to a large extent it was an incidental impact, the result of the existence of a new social and political entity rather than of a doctrine. Earliest Islām as seen in the Qurʾān or in the more verifiable accounts of the Prophet's life simply do not deal with the arts, either on the practical level of requiring or suggesting forms as expressions of the culture or on the ideological level of defining a Muslim attitude toward images. In all instances, concrete Qurʾānic passages later used for the arts had their visual significance extrapolated.

There is no prohibition against representations of living things, and not a single Qurʾānic passage refers clearly to the mosque, eventually to become the most characteristically Muslim religious building. In the simple, practical, and puritanical milieu of early Islām, aesthetic or visual questions simply did not arise.

Influence of Islām in establishing a new artistic tradition

The impact of the faith on the arts occurred rather as the fledgling culture encountered the earlier non-Islāmic world and sought to justify its own acceptance or rejection of new ways and attitudes. The discussion oi two examples of particular significance illustrates the point. One is the case of the mosque. The word itself derives from the Arabic *masjid,* "a place where one prostrates one's self (in front of God)." It was a common term in pre-Islāmic Arabic and in the Qurʾān, where it is applied to sanctuaries in general without restriction. If a more concrete significance was meant, the word was used in construct with some other term, as in *masjid al-ḥaram* to refer to the Meccan sanctuary. There was no need in earliest times for a uniquely Muslim building, for any place could be used for private prayer as long as the correct direction (*qiblah,* originally Jerusalem, but very soon Mecca) was observed and the proper sequence of gestures and pious statements was followed. In addition to private prayer, which had no formal setting, Islām instituted a collective prayer on Fridays, where the same ritual was accompanied by a sermon from the *imām* (leader of prayer, originally the Prophet, then his successors, and later legally any able-bodied Muslim) and by the more complex ceremony of the *khuṭbah,* a collective swearing of allegiance to the community's leadership. This ceremony served to strengthen the common bond between all members of the *ummah,* the Muslim "collectivity," and its importance in creating and maintaining the unity of early Islām has often been emphasized. There were two traditional locales for this event in the Prophet's time. One was his private house, whose descriptions have been preserved; it was a large open space with private rooms on one side and rows of palm trunks making a colonnade on two other sides, the deeper colonnade being on the side of the *qiblah.* The Prophet's house was not a sanctuary but simply the most convenient place for the early community to gather. Far less is known about the second place of gathering for the Muslim community. It was used primarily on major feast days, such as the end of the fasting period or the feast of sacrifice. It was called a *muṣallā,* literally "a place for prayer," and *muṣallā*s were usually located outside city walls. Nothing is known about the shape taken by *muṣallā*s, but in all probability they were as simple as pre-Islāmic pagan sanctuaries: large enclosures surrounded by a wall and devoid of any architectural or ornamental feature.

Altogether then there was hardly anything that could be identified as a holy building or as an architectural form. To be complete, one should add two additional features. One is an action, the call to prayer (*adhān*). It became, fairly rapidly, a formal moment preceding the gathering of the faithful. One man would climb on the roof and proclaim that God is great and that men must congregate to pray. There was no formal monument attached to the ceremony, though it led eventually to the ubiquitous minaret. The other early feature was an actual structure. It was the *minbar,* a chair with several steps on which the Prophet would climb in order to preach. The monument itself had a pre-Islāmic origin, but Muḥammad transformed it into a characteristically Muslim form.

With the exception of the *minbar,* only a series of actions was formulated in early Islāmic times. There were no forms attached to them, nor were any needed. But, as the Muslim world grew in size, the contact with many other cultures brought about two developments. On the one hand there were thousands of examples of beautiful religious buildings that impressed the conquering Arabs. But, more importantly, the need arose to preserve the restricted uniqueness of the community of faithful and to express its separateness from other groups. Islāmic religious architecture began with this need and, in ways to be described later, created a formal setting for the activities, ceremonies, and ideas that had been formless at the outset.

Muslim iconoclasm

A second and closely parallel development of the impact of the Islāmic religion on the visual arts is the celebrated question of a Muslim iconoclasm. As has already been mentioned, the Qur'ān does not utter a word for or against the representation of living things. It is equally true that from about the middle of the 8th century a prohibition had been formally stated, and thenceforth it would be a standard feature of Islāmic thought, even though the form in which it is expressed has varied from absolute to partial and even though it has never been totally followed. The justification for the prohibition tended to be that any representation of a living thing was an act of competition with God, for he alone can create something that is alive. It is striking that this theological explanation reflects the state of the arts in the Christian world at the time of the Muslim conquest—a period of iconoclastic controversy. It may thus be suggested that Islām developed an attitude toward images as it came into contact with other cultures and that its attitude was negative because the arts of the time appeared to lead easily to dreaded idolatry. While it is only by the middle of the 8th century that there is actual proof of the existence of a Muslim doctrine, it is likely that, more or less intuitively, the Muslims felt a certain reluctance toward representations from the very beginning. For all monuments of religious art are devoid of any representations; even a number of attempts at representational symbolism in the official art of coinage were soon abandoned.

This rapid crystallization of Islāmic attitudes toward images has considerable significance. For practical purposes, representations are not found in religious art, although matters are quite different in secular art. Instead there occurred very soon a replacement of imagery with calligraphy and the concomitant transformation of calligraphy into a major artistic medium. Furthermore, the world of Islām tended to seek means of representing the holy other than by images of men, and one of the main problems of interpretation of Islāmic art is that of the degree of means it achieved in this search. But there is a deeper aspect to this rejection of holy images. Although the generally Semitic or specifically Jewish sources that have been given to Islāmic iconoclasms have probably been exaggerated, the reluctance imposed by the circumstances of the 7th century transformed into a major key of artistic creativity the magical fear of visual imagery that exists in all cultures but that is usually relegated to a secondary level. This uniqueness is certainly one of the main causes of the abstract tendencies that are among the great glories of the tradition. Even when a major art of painting did develop, it remained always somehow secondary to the mainstream of the culture's development.

Both in the case of the religious building and in that of the representations, therefore, it was the contact with pre-Islāmic cultures in Muslim-conquered areas that compelled Islām to transform its practical and unique needs into monuments and to seek within itself for intellectual and theological justifications for its own instincts. The great strength of early Islām was that it possessed within itself the ideological means to put together a visual expression of its own, even though it did not develop at the very beginning a need for such an expression.

One last point can be made about the origins of Islāmic art. It concerns the degree of importance taken by the various artistic and cultural entities conquered by the Arabs in the 7th and 8th centuries, for the early empire had gathered in regions that had not been politically or even ideologically related for centuries. During the first century or two of Islām, the main models and the main sources of inspiration were certainly the Christian centres around the Mediterranean. But the failure to capture Constantinople and to destroy the Byzantine Empire also made these Christian centres inimical competitors, whereas the whole world of Iran became an integral part of the empire, even though the conquering Arabs were far less familiar with the latter than with the former. A much more complex problem is posed by conversions, for it is through the success of the militant Muslim religious mission that the culture expanded so rapidly. Insofar as one can judge, it is the common folk, primarily in cities, who took over the new faith most rapidly; and thus there was added in early Islāmic culture a folk element whose impact may have been larger than has hitherto been imagined.

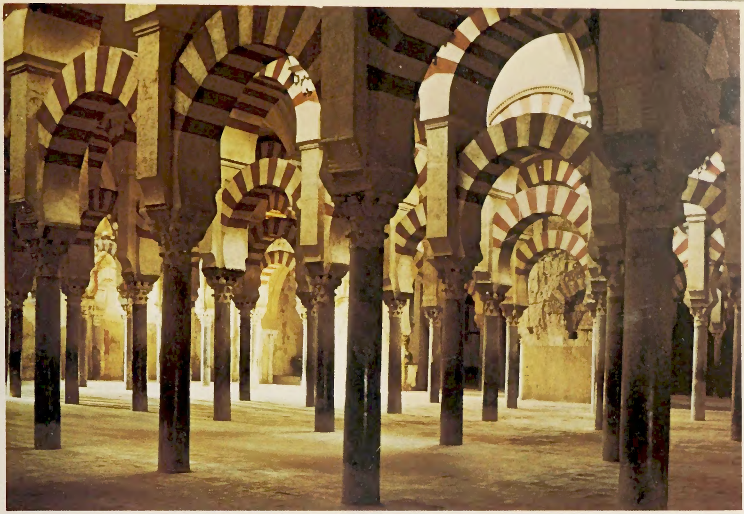
Folk element in Islāmic art

These preliminary considerations on the origins of Islāmic art have made it possible to outline several of the themes and problems that remained constant features of the tradition: a self-conscious sense of uniqueness when compared to others; a continuous reference to its own Qur'ānic sources; a constant relationship to many different cultures; a folk element; and a variety of regional developments. None of these features remained constant, not even those aspects of the faith that affected the arts. But while they changed, the fact of their existence, their structural presence, remained a constant of Islāmic art.

EARLY PERIOD: THE UMAYYAD AND 'ABBĀSID DYNASTIES

Of all the recognizable periods of Islāmic art, this is by far the most difficult one to explain properly, even though it is quite well documented. There are two reasons for this difficulty. On the one hand, it was a formative period, a time when new forms were created that identify the aesthetic and practical ideals of the new culture. Such periods are difficult to define when, as in the case of Islām, there was no artistic need inherent to the culture itself. The second complication derives from the fact that Muslim conquest hardly ever destroyed former civilizations with its own established creativity. Material culture, therefore, continued as before, and archaeologically it is almost impossible to distinguish between pre-Islāmic and early Islāmic artifacts. Paradoxical though it may sound, there is an early Islāmic Christian art of Syria and Egypt, and in many other regions the parallel existence of a Muslim and of a non-Muslim art continued for centuries. What did happen during early Islāmic times, however, was the establishment of a dominant new taste, and it is the nature and character of this taste that has to be explained. It occurred first in Syria and Iraq, the two areas with the largest influx of Muslims and with the two successive capitals of the empire, Damascus under the Umayyads and Baghdad under the early 'Abbāsids. From Syria and Iraq this new taste spread in all directions and adapted itself to local conditions and local materials, thus creating considerable regional and chronological variations in early Islāmic art.

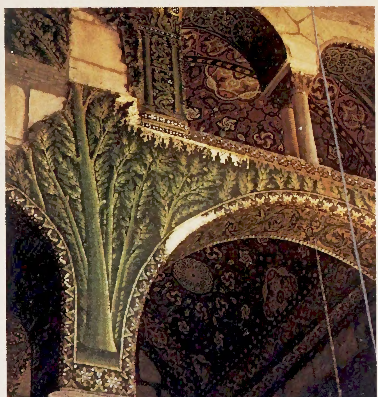
From a historical point of view two major dynasties are involved. One is the Umayyad dynasty, which ruled from 661 to 750 and whose monuments are datable from 680 to 745. It was the only Muslim dynasty ever to control the whole of the Islāmic-conquered world. The second dynasty is the 'Abbāsid dynasty; technically its rule extended as late as 1258, but in reality its princes ceased to be a significant cultural factor after the second decade of the 10th century. The 'Abbāsids no longer controlled Spain, where an independent Umayyad caliphate had been established; and in Egypt as well as in northeastern Iran a number of more or less independent dynasties appeared, such as the Ṭūlūnids or the Sāmānids. Although recent research tends to make the conclusion less certain than it used to be for the Sāmānids and northeast Iran, the initial impulse for the artistic creativity of these dynasties came from the main 'Abbāsid centres in Iraq. While in detailed studies it is possible to distinguish between Umayyad and 'Abbāsid art or between the arts of various provinces, the key features of the first three centuries of Islāmic art (roughly through the middle of the 10th century) are the interplay between local or imperial impulses and the creation of new forms and functions.



Interior of the Great Mosque of Córdoba, Spain, begun 785. The building is now a Christian cathedral.

Umayyad and ʿAbbasid Art

Bowl from Nīshāpūr, lead-glazed earthenware with a slip decoration. In the Victoria and Albert Museum, London.



Mosaics decorating the portico of the Great Mosque of Damascus, Syria, 715.

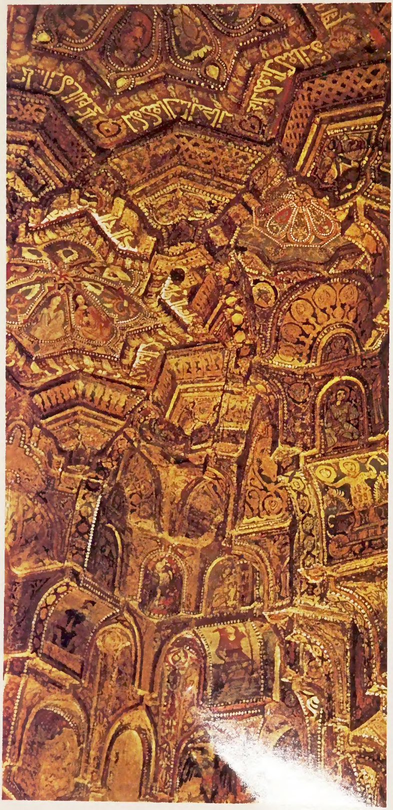
Woven silk bearing the inscription "Glory and happiness to Qaid Abul Mansur Nudjkatin; may God continue his prosperity," 10th century. In the Louvre, Paris. 94 × 52 cm.

Plate 1: By courtesy of (centre left) the Victoria and Albert Museum, London; photographs, (top) Alfonso Gutierrez Escera—Ostman Agency, (bottom left) Cliche Musees Nationaux, (bottom right) J.E. Dayton, London

Fatimid art

Coronation mantle of King Roger II of Sicily, 1133. Gold embroidery and pearls on a red silk ground. In the Hofburg, Vienna.

Ceiling of the Capella Palatina, Palermo, Sicily. The chapel was built by the Norman kings of Sicily and decorated by Fātimid artists.



Bowl of lustre-ware by the potter Sa'ad, depicting a Christian priest swinging a censer, first half of the 12th century. In the Victoria and Albert Museum, London.

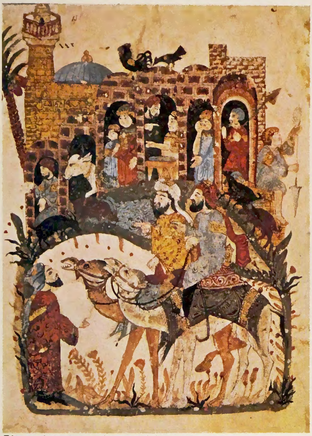
Bronze griffin. In the Camposanto, Pisa, Italy.

Plate 2: By courtesy of (top) the Hofburg Kunsthistorisches Museum, Vienna, (centre right) the Victoria and Albert Museum, London; photographs, (left) M. Desjardins—Realities, (bottom right) SCALA—Art Resource/EB Inc.

Plate 3: By courtesy of (top right) the Topkapı Saray Museum, Istanbul, (centre left) the Bibliotheque Nationale, Paris, (bottom left) the Smithsonian Institution, Freer Gallery of Art, Washington, D.C.; photograph, (right) Roland and Sabrina Michaud—Rapho/Photo Researchers from EB Inc.

Seljuq art

"Golshāh has removed her veil during a battle," miniature from *Varqeh o-Golshāh*, 13th century. In the Topkapı Saray Museum, Istanbul. (Ms. Hazine 841, fol. 22.) 10.2 × 29.7 cm.



Discussion near a village, miniature painted by Yahyā ibn Mahmūd al-Wāsitī from the 43rd *maqāmah* of the *Maqāmāt* ("Assemblies") of al-Harīrī, 1237. In the Bibliothèque Nationale, Paris. (Ms. Arabe 5847, folio 138 r.) 34.8 × 26 cm.

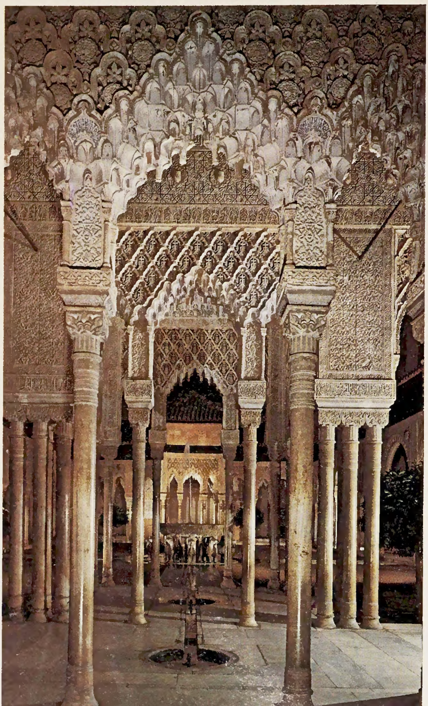
Lustre dish depicting Khosrow II as he discovers Shīrīn bathing, by Sayyid Shams ad-Dīn al Husani, from Kāshān, Iran, c. 1210. In the Freer Gallery of Art, Washington, D.C.

The minaret of Jām, Afghanistan, 1116–1202.

Ivory casket, 13th century. In the Palazzo Reale, Capella Palatina, Palermo, Sicily. 39 × 40 × 24 cm.

Al-Hārith talks to Abū Zayd in his tent, miniature from the 26th *maqāmah* of the *Maqāmāt* ("Assemblies") of al-Harīrī, probably Egyptian, 1334. In the Österreichische Nationalbibliothek, Vienna (MS. A. F. 9, folio 87 v). Miniature only, 13.7 × 15.8 cm.

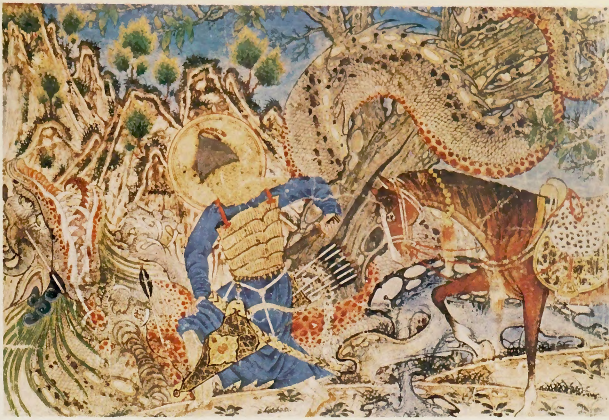
Moorish and Mamluk art



Court of the Lions, the Alhambra, Granada, Spain, 14th century.

Hanging mosque lamp, enamelled and gilded glass, from Aleppo, Syria, *c.* 1300. In the Museum für Islamische Kunst, Staatliche Museen Preussischer Kulturbesitz, West Berlin.

Plate 4: By courtesy of (top left) the Soprintendenza alle Gallerie ed Alle Opere d'Arte della Sicilia, Palermo, (top right) the Osterreichische Nationalbibliothek, Vienna, (bottom right) the Museum fur Islamische Kunst, Staatliche Museen Preussischer Kulturbesitz, Berlin—EB Inc.; photograph, (bottom left) Raffaello Bencini—SCALA/Art Resource from EB Inc.



Bahrām Gūr killing a dragon, illustration from the *Shāh-nāmeh* ("Book of Kings") of Ferdowsī, known as the Demotte *Shāh-nāmeh*, 1320–60, from Tabriz, Iran. In the Cleveland Museum of Art. Height 40.6 cm.

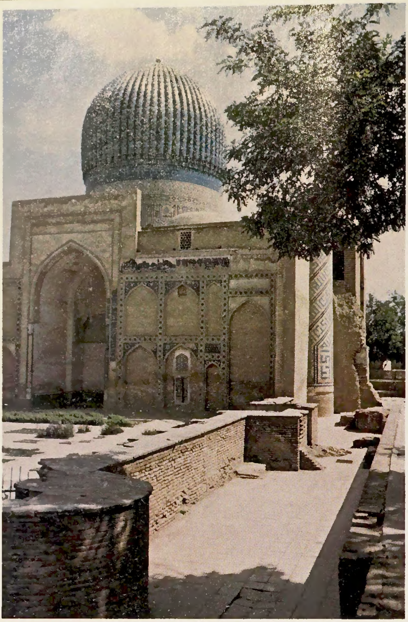
Il-Khanid art of the Mongol Period

Pottery bowl from Kāshān, Iran, late 14th century. In the Victoria and Albert Museum, London.

Dīwān of Sultan Ahmad, pastoral border painted by Junayd, *c.* 1405, from Baghdad. In the Freer Gallery of Art, Washington, D.C. 29.2 × 20.3 cm.

Mongol warriors, miniature from Rashīd ad-Dīn's *History of the World*, 1307. In the Edinburgh University Library, Scotland. Miniature only, 25 × 11.4 cm.

Plate 5 By courtesy of (top) the Cleveland Museum of Art, Ohio, Grace Rainey Rogers Fund, (centre left) the Victoria and Albert Museum, London, (bottom left) the Edinburgh University Library, Scotland, (bottom right) the Smithsonian Institution, Freer Gallery of Art, Washington, D.C.; photograph (centre left) John Webb



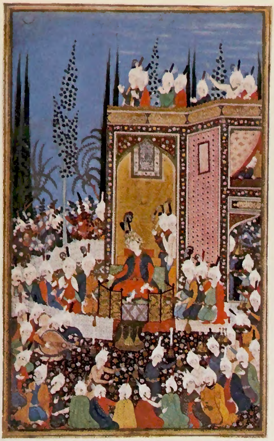
The mausoleum of Timur at Samarkand, 1434.

"Prince Humāy at the Gate of Humāyun's Castle," miniature painted by Junayd for the *Khamseh* of Khwāju Kermānī, 1396. In the British Library (MS. Add 18113, folio 18v). 29 × 20.2 cm.

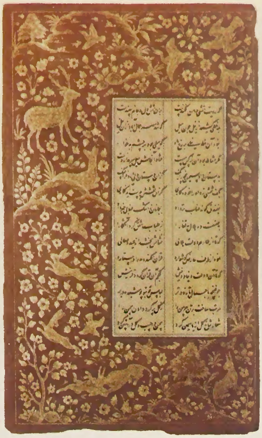
Timurid art of the Mongol Period

Capture of the fortress of the Knights Hospitallers at Smyrna, miniature from a *Zafar-nāmeh* (a life of Timur) by Behzād, *c.* 1490, from Herāt. In the John Work Garrett Library, Johns Hopkins University, Baltimore. 25.2 × 13 cm.

Section of relief tilework from the mausoleum of Bayram Khān at Fatḥābād, Uzbekistan, late 14th to early 15th century. In the Victoria and Albert Museum, London. Length 1.52 m.

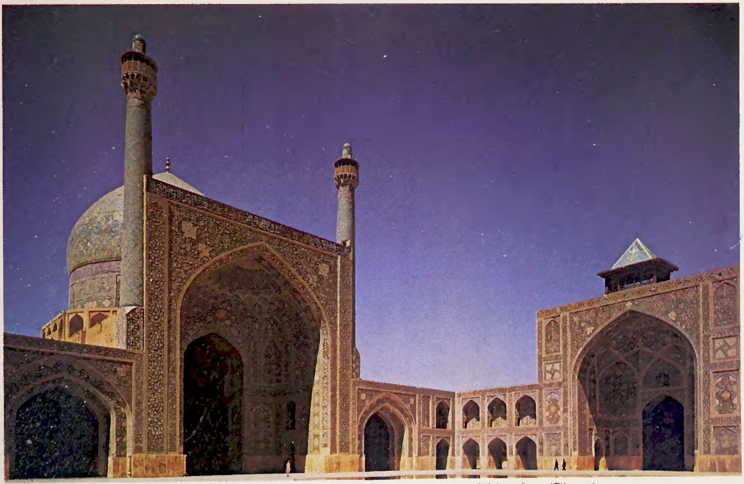


The Feast of 'Id, illustration from a *Divān* of Hāfez, signed Sultan Muhammad, *c.* 1520. In a private collection. 24 × 16 cm.



Miniature from *Yusof o-Zalikhā* by Jāmī, the text in small *nasta 'liq* calligraphy, 1557. In the Freer Gallery of Art, Washington, D.C. 25.2 × 15 cm.

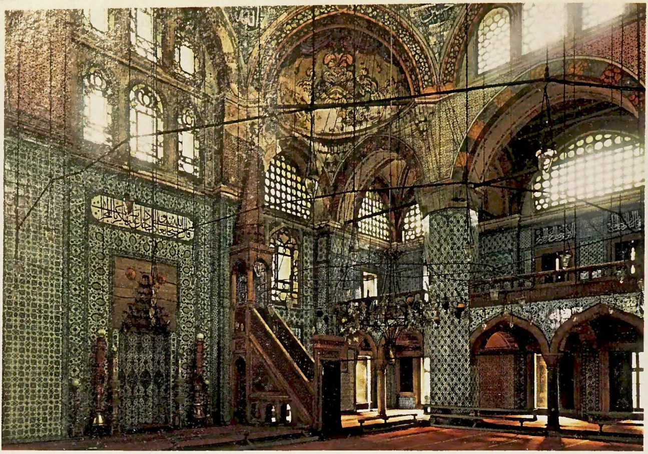
Safavid art



Two *eyvān*s of the Masjed-e Shāh of 'Abbās I the Great at Isfahan, Iran, 17th century.

Plate 6: By courtesy of (centre right) the John Work Garrett Library, Johns Hopkins University, Baltimore, (bottom) Victoria and Albert Museum, London; photographs, (top right) reproduced by permission of the British Library, (top left) Alex Langley—Photo Researchers/EB Inc., (centre right) Walters Art Gallery, Baltimore, (bottom) John Webb

Plate 7: By courtesy of (top left) a private collector, Cambridge, Mass. (formerly in the collection of Louis J. Cartier), (top right) the Smithsonian Institution, Freer Gallery of Art, Washington, D.C.; photographs, (top left) Fogg Art Museum, Cambridge, Mass., (bottom) Ray Manley—Shostal/EB Inc



Interior of the Rüstem Paşa Mosque, Istanbul, showing the coloured tile decoration.

Ottoman art

The Sultan watching dancers and comedians in the Hippodrome, illustration from the *Surname-i Vehbi* of Ahmed III (1703–30), painted by Levnî. In the Topkapı Saray Museum, Istanbul.

Silk caftan said to be that of Ba... (1481–1512). In the Topkapı Sara... Istanbul.

Isnik ware dish, second half of the 16th century. In the Victoria and Albert ... London. Diameter 30.5 cm.

"Star Ushak" carpet from western Anatolia, late 16th to early 17th century. In the Metropolitan Museum of Art, New York City. 2.17 × 4.27 m.

Plate 8: By courtesy of (bottom left, centre) the Topkapı Sa... ...entre) the Victoria and Albert Museum, London, (bottom right) the Metropolitan Museum of Art, New York, gift of Joseph V. McMullen; photographs, (top, bo... ...Inc., (bottom centre) John Webb, (bottom right) Otto E. Nelson

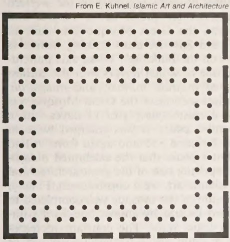
It is possible to study these centuries as a succession of clusters of monuments, but, since there are so many of them, a study can easily end up as an endless list. It is preferable, therefore, to centre the discussion of Umayyad and ʿAbbāsid monuments on the functional and morphological characteristics that identify the new Muslim world and only secondarily be concerned with stylistic progression or regional differences.

Origin of the mosque

Architecture. *Religious buildings.* The one obviously new function developed during this period is that of the mosque, or *masjid.* The earliest adherents of Islām used the private house of the Prophet in Medina as the main place for their religious and other activities and *muṣallā*s without established forms for certain holy ceremonies. The key phenomenon of the first decades that followed the conquest is the creation outside of Arabia of *masjid*s in every centre taken over by the new faith. These were not simply or even primarily religious centres. They were rather the community centres of the faithful, in which all social, political, educational, and individual affairs were transacted. Among these activities were common prayer and the ceremony of the *khuṭbah.* The first mosques were built primarily to serve as the restricted space in which the new community would take its own collective decisions. It is there that the treasury of the community was kept, and early accounts are full of anecdotes about the immense variety of events, from the dramatic to the scabrous, that took place in mosques. Since even in earliest times the Muslim community consisted of several superimposed and interconnected social systems, mosques reflected this complexity, and, next to large mosques for the whole community, tribal mosques and mosques for various quarters of a town or city are also known.

None of these early mosques has survived, and no descriptions of the smaller ones have been preserved. There do remain, however, accurate textual descriptions of the large congregational buildings erected at Kūfah and Basra in Iraq and at al-Fusṭāṭ in Egypt. At Kūfah a larger square was marked out by a ditch, and a covered colonnade known as a *ẓullah* (a shady place) was put up on the *qiblah* side. In 670 a wall pierced by many doors was built in place of the ditch, and colonnades were put up on all four sides, with a deeper one on the *qiblah.* In all probability the Basra mosque was very similar, and only minor differences distinguished the ʿAmr ibn al-ʿĀṣ mosque at al-Fusṭāṭ. Much has been written about the sources of this type of building, but the simplest explanation may be that this is the very rare instance of the actual creation of a new architectural type. The new faith's requirement for centralization, or a space for a large and constantly growing community, could not be met by any existing architectural form. Almost accidentally, therefore, the new Muslim cities of Iraq created the hypostyle mosque (a building with the roof resting on rows of columns). A flexible architectural unit, a hypostyle structure could be square or rectangular and could be increased or diminished in size by the addition or subtraction of columns. The single religious or symbolic feature of the hypostyle mosque was a *minbar* (a pulpit) for the preacher, and the direction of prayer was indicated by the greater depth of the colonnade on one side of the structure.

From E. Kuhnel, *Islamic Art and Architecture*



Plan of reconstructed mosque at Kūfah.

Versatility of the hypostyle plan

The examples of Kūfah, Basra, and al-Fusṭāṭ are particularly clear because they were all built in newly created cities. Matters are somewhat more complex when discussing the older urban centres taken over by Muslims. Although it is not possible to generalize with any degree of certainty, two patterns seem to emerge. In some cases, such as Jerusalem and Damascus and perhaps in most cities conquered through formal treaties, the Muslims took for themselves an available unused space and erected on it some shelter, usually a very primitive one. In Jerusalem this space happened to be a particularly holy one—the area of the Jewish Temple built by Herod I the Great, which had been left willfully abandoned and ruined by the triumphant Christian empire. In Damascus it was a section of a huge Roman temple area, on another part of which there was a church dedicated to St. John the Baptist. Unfortunately too little is known about other cities to be able to demonstrate that this pattern was a common one. The very same uncertainty surrounds the second pattern, which consisted in forcibly transforming sanctuaries of older faiths into Muslim ones. This was the case at Ḥamāh in Syria and at Yazd-e Khvāst in Iran, where archaeological proof exists of the change. There are also several literary references to the fact that Christian churches, Zoroastrian fire temples, and other older abandoned sanctuaries were transformed into mosques. Altogether, however, these instances probably were not too numerous, because in most places the Muslim conquerors were quite anxious to preserve local tradition and because few older sanctuaries could easily serve the primary Muslim need of a large centralizing space.

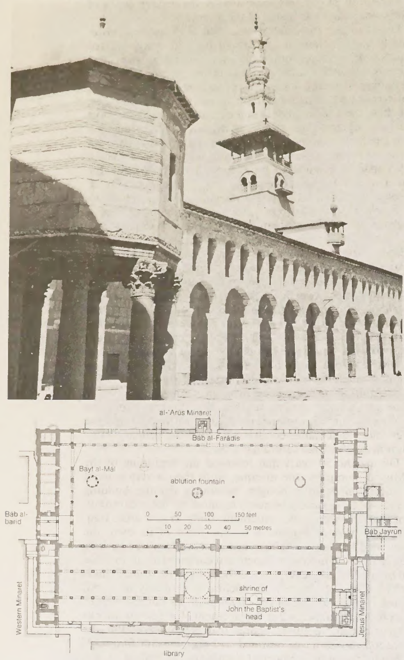
Transformation of existing pre-Islāmic sanctuaries into mosques

During the 50 years that followed the beginning of the Muslim conquest, the mosque, until then a very general concept in Islāmic thought, became a definite building reserved for a variety of needs required by the community of faithful in any one settlement. Only in one area, Iraq, did the mosque acquire a unique form of its own, the oriented hypostyle. Neither in Iraq nor elsewhere is there evidence of symbolic or functional components in mosque design. The only exception is that of the *maqṣūrah* (literally "closed-off space"), an enclosure, probably in wood, built near the centre of the *qiblah* wall. Its purpose was to protect the caliph or his replacement, for several attacks against major political figures had taken place. But the *maqṣūrah* was never destined to be a constant fixture of mosques, and its typological significance is limited.

During the rule of the Umayyad prince al-Walīd I (705–715), a number of complex developments within the Muslim community were crystallized in the construction of three major mosques, at Medina, Jerusalem, and Damascus. The very choice of these three cities is indicative: the city in which the Muslim state was formed and in which the Prophet was buried; the city held in common holiness by Jews, Christians, and Muslims, to which was rapidly accruing the mystical hagiography surrounding the Prophet's ascension into heaven; and the ancient city that became the capital of the new Islāmic empire. A first and essential component of al-Walīd's mosques was, thus, their imperial character; they were to symbolize the permanent establishment of the new faith and of the state that derived from it. They were no longer purely practical shelters but willful monuments.

Mosques of Medina, Jerusalem, and Damascus

Although the plans of al-Aqṣā Mosque in Jerusalem and of the mosque of Medina can be reconstructed with a fair degree of certainty, only the one at Damascus has been preserved with comparatively minor alterations and repairs. In plan the three buildings appear at first glance to be quite different from each other. The Medina mosque was essentially a large hypostyle with a courtyard. The colonnades on all four sides were of varying depth. Al-Aqṣā Mosque consisted of an undetermined number of naves (possibly as many as 15) parallel to each other in a north–south direction. There was no courtyard because the rest of the huge esplanade of the former Jewish temple served as the open space in front of the building. The Umayyad Mosque of Damascus is a rectangle 515 by 330 feet (157 by 100 metres) whose outer limits and three gates



Great Mosque of Damascus, Syria, built by al-Walīd I, 705–715. (Top) Courtyard with the Bayt al-Māl (treasury) on the left, beyond which can be seen one of the three towers that were the first minarets in Islām. (Below) Plan.
Paul Almasy

are parts of a Roman temple (a fourth Roman gate on the *qiblah* side was blocked). The interior consists of an open space surrounded on three sides by a portico and of a covered space of three equal long naves parallel to the *qiblah* wall that are cut in the middle by a perpendicular nave.

The three buildings share several important characteristics. They are all large spaces with a multiplicity of internal supports; and although only the Medina mosque is a pure hypostyle, the Jerusalem and Damascus mosques have the flexibility and easy internal communication characteristic of a hypostyle building. All three mosques exhibit a number of distinctive new practical elements and symbolic meanings. Many of these occur in all mosques; others are only known in some of them. The *miḥrāb*, for example, appears in all mosques. This is a niche of varying size that tends to be heavily decorated. It occurs in the *qiblah* wall, and, in all probability, its purpose was to commemorate the symbolic presence of the Prophet as the first *imām*, although there are other explanations. It is in Damascus only that the ancient towers of the Roman building were first used as minarets to call the faithful to prayer and to indicate from afar the presence of Islām (initially minarets tended to exist only in predominantly non-Muslim cities). All three mosques are also provided with an axial nave, a wider aisle unit on the axis of the building, which served both as a formal axis for compositional purposes and as a ceremonial one for the prince's retinue. Finally, all three buildings were heavily decorated with marble, mosaics, and woodwork. At least in the mosque of Damascus, it is further apparent that there was careful concern for the formal composition—a balance between parts that truly makes this mosque a work of art. This is particularly evident in the successful relationship established between the open space of the court and the facade of the covered *qiblah* side.

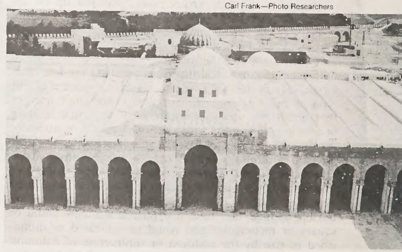
Appearance of the *miḥrāb* and minaret

When compared to the first Muslim buildings of Iraq and Egypt, the monuments of al-Walīd are characterized by the growing complexity of their forms, by the appearance of uniquely Muslim symbolic and functional features, and by the quality of their construction. While the dimensions, external appearance, and proportions of any one of them were affected in each case by unique local circumstances, the internal balance between open and covered areas and the multiplicity of simple and flexible supports indicate the permanence of the early hypostyle tradition.

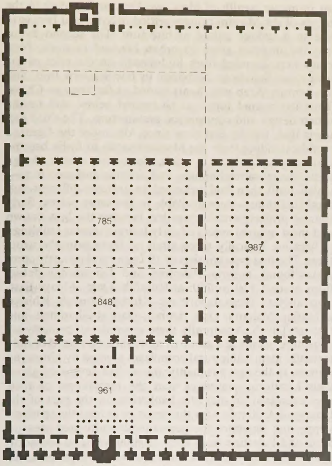
Either in its simplest form, as in Medina, or in its more formalized shape, as in Damascus, the hypostyle tradition dominated mosque architecture from 715 to the 10th century. As it occurs at Nīshāpūr in northeastern Iran, Sīrāf in southern Iran, al-Qayrawān (Kairouan) in Tunisia, and Córdoba in Spain, it can indeed be considered as the classic early Islāmic type. Its masterpieces occur in Iraq and in the West. The monumentalization of the early Iraqi hypostyle is illustrated by the two ruined structures in Sāmarrā', with their enormous sizes (790 by 510 feet [240 by 156 metres] for one and 700 by 440 feet [213 by 135 metres] for the other), their multiple entrances, their complex piers, and, in one instance, a striking separation of the *qiblah* area from the rest of the building. The best preserved example of this type is the mosque of Ibn Ṭūlūn at Cairo (876–879), where a semi-independent governor, Aḥmad ibn Ṭūlūn, introduced Iraqi techniques and succeeded in creating a masterpiece of composition.

Two classic examples of early mosques in the western Islāmic world of interest are preserved in Tunisia and Spain. In al-Qayrawān the Great Mosque was built in stages between 836 and 866. Its most striking feature is the formal emphasis on the building's T-like axis punctuated by two domes, one of which hovers over the earliest preserved ensemble of *miḥrāb*, *minbar*, and *maqṣūrah*. At Córdoba the earliest section of the Great Mosque was built in 785–786. It consisted simply of 11 naves with a wider central one and a court. It was enlarged twice in length, first between 833 and 855 and again from 961 to 965 (it was in the latter phase that the celebrated *maqṣūrah* and *miḥrāb*, comprising one of the great architectural ensembles of early Islāmic art, were constructed). Finally, in 987–988 an extension of the mosque was completed to the east that increased its size by almost one-third without destroying its stylistic unity. The constant increases in the size of this mosque are a further illustration of the flexibility of the hypostyle and its adaptability to any spatial requirement. The most memorable aspects of the Córdoba mosque, however, lie in its construction and decoration. The particularly extensive and heavily decorated

Carl Frank—Photo Researchers



The Great Mosque in al-Qayrawān, Tunisia, built between 836 and 866.



Plan of the Great Mosque at Córdoba, Spain, showing dates of different additions.
From G. Marcais, *L'Architecture Musulmane D'Occident*

miḥrāb area exemplifies a development that started with the Medina mosque and would continue: an emphasis on the *qiblah* wall.

Although the hypostyle mosque was the dominant plan, it was not the only one. From very early Islāmic times, a fairly large number of aberrant plans also occur. Most of them were built in smaller urban locations or were secondary mosques in larger Muslim cities. It is rather difficult, therefore, to evaluate whether their significance was purely local or whether they were important for the tradition as a whole. Since a simple type of square subdivided by four piers into nine-domed units occurs at Balkh in Afghanistan, at Cairo, and at Toledo, it may be considered a pan-Islāmic type. Other types, a single square hall surrounded by an ambulatory, or a single long barrel-vault parallel or perpendicular to the *qiblah*, are rarer and should perhaps be considered as purely local. These are particularly numerous in Iran, where it does seem that the mainstream of early Islāmic architecture did not penetrate very deeply. Unfortunately, the archaeological exploration of Iran is still in its infancy, and many of the mud-brick buildings from the early Islāmic period have been destroyed or rebuilt beyond recognition. As a result, it is extremely difficult to determine the historical importance of monuments found at Neyrīz, Moḥammadīyeh (near Nā'īn), Fahraj (near Yazd), or Hazareh (near Samarkand). For an understanding of the mosque's development and of the general dynamics of Islāmic architecture however, an awareness of these secondary types, which may have existed outside of Iran as well, is essential.

Early aberrant plans

The function of the mosque, the central gathering place of the Muslim community, became the major and most original completely Muslim architectural effort. The mosque was not a purely religious building, at least not at the beginning; but, because it was restricted to Muslims, it is appropriate to consider it as such. This, however, was not the only type of early Islāmic building to be uniquely Muslim. Three other types can be defined architecturally, and a fourth one only functionally.

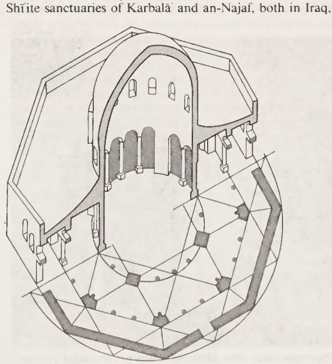
The first type, the Dome of the Rock in Jerusalem, is a unique building. Completed in 691, this masterwork of Islāmic architecture is the earliest major Islāmic monument. Its octagonal plan, use of a high dome, and building techniques are hardly original, although its decoration is unique. Its purpose, however, is what is most remarkable about the building. Since the middle of the 8th century, the Dome of the Rock has become the focal centre of the most mystical event in the life of the Prophet: his ascension into heaven from the rock around which the building was erected. According to an inscription preserved since the erection of the dome, however, it would seem that the building did not originally commemorate the Prophet's ascension but rather the Christology of Islām and its relationship to Judaism. It seems preferable, therefore, to interpret the Dome of the Rock as a victory monument of the new faith's ideological and religious claim on a holy city and on all the religious traditions attached to it.

The *ribāṭs* of the frontier

The second distinctly Islāmic type of religious building is the little-known *ribāṭ*. As early as in the 8th century, the Muslim empire entrusted the protection of its frontiers, especially the remote ones, to warriors for the faith (*murābiṭūn*, "bound ones") who lived, permanently or temporarily, in special institutions known as *ribāṭs*. Evidence for these exist in Central Asia, Anatolia, and North Africa. It is only in Tunisia that *ribāṭs* have been preserved. The best one is at Sūsah, Tunisia; it consists of a square fortified building with a single fairly elaborate entrance and a central courtyard. It has two stories of private or communal rooms. Except for the prominence taken by an oratory, this building could be classified as a type of Muslim secular architecture. Since no later example of a *ribāṭ* is known, there is some uncertainty as to whether the institution ever acquired a unique architectural form of its own.

The diverse functions of the Muslim mausoleum

The last type of religious building to develop before the end of the 10th century is the mausoleum. Originally Islām was strongly opposed to any formal commemoration of the dead. But three independent factors slowly modified an attitude that was eventually maintained only in the most strictly orthodox circles. One factor was the growth of the Shī'ite heterodoxy, which led to an actual cult of the descendants of the Prophet through his son-in-law 'Alī. The second factor was that, as Islām strengthened its hold on conquered lands, a wide variety of local cultic practices and especially the worship of certain sacred places began to affect the Muslims, resulting in a whole movement of Islāmization of ancient holy places by associating them with deceased Muslim heroes and holy men or with prophets. The third factor is not, strictly speaking, religious, but it played a major part. As more or less independent local dynasties began to grow, they sought to commemorate themselves through mausoleums. Not many mausoleums have remained from these early centuries, but literary evidence is clear on the fact that the Shī'ite sanctuaries of Karbalā' and an-Najaf, both in Iraq,



Cross section of the Dome of the Rock, Jerusalem.

and Qom, Iran, already possessed monumental tombs. At Sāmarrā' an octagonal mausoleum had been built for three caliphs. The masterpieces of early funerary architecture occur in Central Asia, such as the royal mausoleum of the Sāmānids (known incorrectly as the mausoleum of Esmā'īl the Sāmānid) at Bukhara (before 942), which is a superb example of Islāmic brickwork. In some instances a quasi-religious character was attached to the mausoleums, such as the one at Tim (976), which already has the high facade typical of so many later monumental tombs. In all instances the Muslims took over or rediscovered the ancient tradition of the centrally planned building as the characteristic commemorative structure.

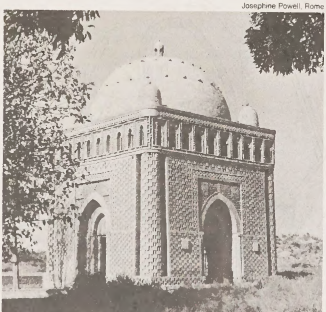
The *madrasah*

The fourth kind of Muslim building is the *madrasah,* an institution for religious training set up independently of mosques. It is known from texts that such privately endowed schools existed in the northeastern Iranian world as early as in the 9th century, but no description exists of how they were planned or looked.

Secular architecture. Whereas the functions of the religious buildings of early Islām could not have existed without the new faith, the functions of secular Muslim architecture have *a priori* no specifically Islāmic character. This is all the more so since one can hardly point to a significant new need or habit that would have been brought from Arabia by the conquering Muslims and since so little was destroyed in the conquered areas. It can be assumed, therefore, that all pre-Islāmic functions such as living, trading, and manufacturing continued in whatever architectural setting they may have had. Only one exception is certain. With the disappearance of Sāsānian kingship, the pre-Islāmic Iranian imperial tradition ceased, and elsewhere conquered minor kings and governors left their palaces and castles. A new imperial power was created, located first in Damascus, then briefly in the northern Syrian town of ar-Ruṣāfah, and eventually in Baghdad and Sāmarrā' in Iraq. New governors and, later, almost independent princes took over provincial capitals, which were sometimes old seats of government and, at other times, were new Muslim centres. In all instances, however, there is no reason to assume that for an architecture of power or of pleasure early Muslims would have felt the need to modify pre-Islāmic traditions. In fact there is much in early Islāmic secular architecture that can be used to illustrate secular arts elsewhere—in Byzantium, for example, or even in the West. If any new political or social entity is to succeed in preserving an identity of its own, however, it must give to its secular needs certain directions and emphases that will eventually establish a unique cultural image. This is what happened in the development of Umayyad and early 'Abbāsid secular architecture.

Three factors contributed to the evolution of a new secular architecture. One was that the accumulation of

Josephine Powell, Rome



Royal mausoleum of the Sāmānids, Bukhara, Uzbek S.S.R., (before 942).

an immense wealth of ideas, workers, and money in the hands of the Muslim princes settled in Syria and Iraq gave rise to a unique palace architecture. The second factor was the impetus given to urban life and to trade. New cities were founded from Sijilmāssah on the edge of the Moroccan Sahara to Nīshāpūr in northeastern Iran, and 9th-century Arab merchants traded as far away as China. Thus the second topic, to be treated below, will be the urban design and commercial architecture. The third factor is that, for the first time since Alexander the Great, a world extending from the Mediterranean to India became culturally unified. As a result, decorative motifs, design ideas, structural techniques, and artisans and architects—which until then had belonged to entirely different cultural traditions—were available in the same places. Early Islāmic princely architecture has become the best known and most original aspect of early Islāmic secular buildings.

The country princely palace complex

There are basically three kinds of these princely structures. The first type consists of 10 large rural princely complexes found in Syria, Palestine, and Transjordan dating from around 710 to 750: ar-Ruṣāfah, Qaṣr al-Ḥayr East, Qaṣr al-Ḥayr West, Jabal Says, Khirbat Minyah, Khirbat al-Mafjar, Mshattā, Qaṣr 'Amrah, Qaṣr al-Kharānah, and Qaṣr aṭ-Ṭūbah. Apparently these examples of princely architecture belong to a group of more than 60 ruined or only textually identifiable rural complexes erected by Umayyad princes. In the past a romantic theory had developed about their locations, suggesting that the remoteness of their sites expressed an atavistic hankering on the part of the Umayyad Arab rulers for the desert or at least the semiarid steppe that separates the permanently cultivated areas of Syria and Palestine from their original home in the north Arabian wilderness. This theory has been disproved, for every one of these has turned out to have been a major agricultural or trade centre, some of which were developed even before the Muslim conquest. Private palaces were built, notably at ar-Ruṣāfah, Qaṣr al-Ḥayr West, Khirbat al-Mafjar, Qaṣr 'Amrah, and Mshattā. These must be considered as early medieval equivalents of the *villae rusticae* so characteristic in the ancient Roman period. Although each of these had a number of idiosyncrasies that were presumably inspired by the needs and desires of its owner, all of these structures tend to share a number of features that can best be illustrated by Khirbat al-Mafjar.

This palace, the richest of them all, contained a residential unit consisting of a square building with an elaborate entrance, a porticoed courtyard, and a number of rooms or halls arranged on two floors. Few of these rooms seem to have any identifiable function, although at Khirbat al-Mafjar a private oratory, a large meeting hall, and an anteroom leading to a cool underground pool have been identified. The main throne room was on the second floor above the entrance. Its plan is not known but probably resembled the preserved throne rooms or reception halls at Qaṣr 'Amrah and Mshattā, which consisted of a three-aisled hall ending in an apse (semicircular or polygonal domed projection) in the manner of a Roman basilica.

Next to an official residence, there usually was a small mosque, generally a miniaturized hypostyle in plan. The most original feature of these establishments was the bath. The bathing area itself is comparatively small, but every bath had its own elaborate entrance and contained a large hall that, at least in the instance of Khirbat al-Mafjar, was heavily decorated and of an unusual shape. It would appear that these halls were for pleasure—places for music, dancing, and probably occasional orgies. In some instances, as at Qaṣr 'Amrah, the same setting may have been used for both pleasure and formal receptions.

These palaces are important illustrations of the luxurious taste and way of life of the new Near Eastern aristocrats, who settled in the countryside and transformed some of it into places of pleasure. This aspect of these establishments is peculiar to the Umayyad dynasty in Syria and Palestine. Outside of this area and period only one comparable structure has been found—at Ukhayḍir in Iraq, which dates from the early 'Abbāsid period. A number of princely residences of the Central Asian or North African countryside are still too little known but appear not to have had the same development. The other important lesson to draw from them is that few of their features are original. All of them derive from the architectural vocabulary of pre-Islāmic times, and it is in the artistic traditions of the Mediterranean world that most of their sources are found, although the Mshattā throne room does have a number of Sāsānian elements. For this reason these palaces should be considered as major examples of pre-Islāmic secular architecture, for as interesting as these monuments are, they are not part of the Islāmic tradition.

The urban palace

A second type of princely architecture—the urban palace—has been preserved only in texts or literary sources, with the exception of the palace at Kūfah in Iraq. Datable from the very end of the 7th century, this example of princely architecture seems to have functioned both as a residence and as the *dār al-imārah,* or centre of government. This dual function is reflected in the use of separate building units and in the absence of much architectural decoration, which suggests that it reflected an austere official taste. Although suggestions concerning the plans used are occasionally encountered in literary sources, this information is not sufficient to define these early urban official buildings of the Muslims. Nothing is known, for instance, about the great Umayyad palace in Damascus aside from the fact that it had a green dome.

Also poorly documented is a development in urban aristocratic buildings that seems to have begun with the ʿAbbāsids during the last decades of the 8th century. This involved the construction of smaller palaces, probably pavilions in the midst of gardens in or around major cities.

The ʿAbbāsid palace-city

The third type of early Islāmic princely architecture is the palace-city. Several of these huge palaces are part of the enormous mass of ruins at Sāmarrāʾ, the temporary ʿAbbāsid capital from 838 to 883. Jawsaq al-Khāqānī, for instance, is a walled architectural complex nearly one mile to a side that in reality is an entire city. It contains a formal succession of large gates and courts leading to a cross-shaped throne room, a group of smaller living units, basins and fountains, and even a racetrack. Too little is known about the architectural details of these huge walled complexes to lead to more than very uncertain hypotheses. Their existence, however, suggests that they were settings for the very elaborate ceremonies developed by the ʿAbbāsid princes, especially when receiving foreign ambassadors. An account, for instance, in Khaṭīb al-Baghdādī's (died 1071) *Taʾrīkh Baghdad* ("History of Baghdad") of the arrival in Baghdad of a Byzantine envoy in 914 illustrates this point. The meeting with the caliph was preceded by a sort of formal presentation intended to impress the ambassador with the Muslim ruler's wealth and power. Treasures were laid down, thousands of soldiers and slaves in rich clothes guarded them, lions roared in the gardens, and on gilded artificial trees mechanical devices made silver birds chirp. The ceremony was a fascinating mixture of a traditional attempt to recreate paradise on earth and a rather vulgar exhibition of wealth that required a huge space, as in the Sāmarrāʾ palaces. Another important aspect of these palace-cities is that they became part of a myth. The walled enclosure in which thousands lived a life unknown to others and into which simple mortals did not penetrate without bringing their own shroud was transformed into legend. It became the mysterious City of Brass of *The Thousand and One Nights,* and it is from its luxurious glory that occasionally a caliph such as Hārūn ar-Rashīd escaped into the "real" world. Even though there is inadequate information on the ʿAbbāsid palace-city, it was clearly a unique early Islāmic creation, and its impact can be detected from Byzantium to Hollywood.

Urban design

Islāmic secular architecture has left considerable information about cities, for systematic urbanization was one of the most characteristic features of early Muslim civilization. It is much too early to draw any sort of conclusion about the actual physical organization of towns, about their subdivisions and their houses, for only at al-Fusṭāṭ (Cairo) and Sīrāf in Iran is the evidence archaeologically clear, and much of it has not yet been properly published. A huge task remains to be done of relating immense amounts of textual material with scraps of archaeological information scattered from Central Asia to Spain, such as the outer walls and impressive gateway preserved at ar-Raqqah in Syria. In general it can be said that there does not seem to have been any idealized master plan for the internal arrangement of an urban site in contradistinction to Hellenistic or Roman towns. Even mosques or palaces were often located eccentrically and not in the middle of the town. Extraordinary attention was paid to water distribution and conservation, as demonstrated by the magnificent 9th-century cisterns in Tunisia, the 9th-century Nilometer (a device to measure the Nile's level) in Cairo, and the elaborate dams, canals, and sluices of Qaṣr al-Ḥayr in Syria. The construction of commercial buildings on a monumental scale occurred. The most spectacular example is the caravansary of Qaṣr al-Ḥayr East, with its magnificent gate.

The concern for palaces and cities that characterized early Islāmic secular architecture shows itself most remarkably in the construction of Baghdad between 762 and 766–767 by the ʿAbbāsid caliph al-Manṣūr. It was a walled round city whose circular shape served to demonstrate Baghdad's symbolic identity as the navel of the universe. A thick ring of residential quarters was separated by four axial, commercial streets entered through spectacular gates. In the centre of the city there was a large open space with a palace, a mosque, and a few administrative buildings. By its size and number of inhabitants, Baghdad was unquestionably a city; however, its plan so strongly emphasized the presence of the caliph that it was also a palace.

Building materials and technology. The early Islāmic period, on the whole, did not innovate much in the realm of building materials and technology but utilized what it had inherited from older traditions. Stone and brick continued to be used around the Mediterranean, while mud brick usually covered with plaster predominated in Iraq and Iran, with a few notable exceptions like Sīrāf, where a masonry of roughly cut stones set in mortar was more common. The most important novelty was the rapid development in Iraq of a baked brick architecture in the late 8th and 9th centuries. Iraqi techniques were later used in Syria at ar-Raqqah and Qaṣr al-Ḥayr East and in Egypt. Iranian brickwork appears at Mshattā in Jordan. The mausoleum of the Sāmānids in Bukhara is the earliest remaining example of the new brick architecture in northeastern Iran. Wood was used consistently but has usually not been very well preserved, except in Palestine and Egypt where climatic (extreme dryness of Egypt), religious (holiness of Jerusalem sanctuaries), or historic (Egypt was never conquered) factors contributed to the continuous upkeep of wooden objects or architectural elements.

As supports for roofs and ceilings, early Islāmic architecture used walls and single supports. Walls were generally continuous, often buttressed with half towers, and rarely (with exceptions in Central Asia) were they articulated or broken by other architectural features. The most common single support was the base–column–capital combination of Mediterranean architecture. Most columns and capitals were either reused from pre-Islāmic buildings or were directly imitated from older models. In the 9th century in Iraq a brick pier was used, a form that spread to Iran and Egypt. Columns and piers were covered with arches. Most often these were semicircular arches; the pointed, or two-centred, arch was known, but it does not seem that its property of reducing the need for heavy supports had been realized. The most extraordinary technical development of arches occurs in the Great Mosque at Córdoba, where, in order to increase the height of the building in an area with only short columns, the architects created two rows of superimposed horseshoe arches. Almost immediately they realized that such a succession of superimposed arches constructed of alternating stone and brick could be modified to create a variety of patterns that would alleviate the inherent monotony of a hypostyle building. A certain ambiguity remains, however, as to whether ornamental effect or structural technology was the predominate concern in the creation of these unique arched columns.

Arches and vaulting

The majority of early Islāmic ceilings were flat. Gabled wooden roofs, however, were erected in the Muslim world west of the Euphrates and simple barrel vaults to the east. Vaulting, either in brick or in stone, was used, especially in secular architecture. Domes were employed frequently

Dome of the *miḥrāb* in the Great Mosque of Córdoba, Spain, *c.* 961.
Fritz Henle—Photo Researchers/EB Inc.

in mosques, consistently in mausoleums, and occasionally in secular buildings. Almost all domes are on squinches (supports carried across corners to act as structural transitions to a dome). Most squinches, as in the al-Qayrawān domes, are classical Greco-Roman niches, which transform the square room into an octagonal opening for the dome. In Córdoba's Great Mosque a complex system of intersecting ribs is encountered, while at Bukhara the squinch is broken into halves by a transverse half arch. The most extraordinary use of the squinch occurs in the mausoleum at Tim, where the surface of this structural device is broken into a series of smaller three-dimensional units rearranged into a sort of pyramidal pattern. This rearrangement is the earliest extant example of *muqarnas,* or stalactite-like decoration that would later be an important element of Islāmic architectural ornamentation. The motif is so awkwardly constructed at Tim that it must have derived from some other source, possibly the ornamental device of using curved stucco panels to cover the corners and upper parts of walls found in Iran at Nīshāpūr.

Architectural decoration. Early Islāmic architecture is most original in its decoration. Mosaics and wall paintings followed the practices of antiquity and were primarily employed in Syria, Palestine, and Spain. Stone sculpture existed, but stucco sculpture, first limited to Iran, spread rapidly throughout the early Islāmic world. Not only were stone or brick walls covered with large panels of stucco sculpture, but this technique was used for sculpture in the round in the Umayyad palaces of Qaṣr al-Ḥayr West and Khirbat al-Mafjar. The latter was a comparatively short-lived technique, although it produced some of the few instances of monumental sculpture anywhere in the early Middle Ages. A variety of techniques borrowed from the industrial arts were used for architectural ornamentation. The *miḥrāb* wall of al-Qayrawān's Great Mosque, for example, was covered with ceramics, while fragments of decorative woodwork have been preserved in Jerusalem and Egypt.

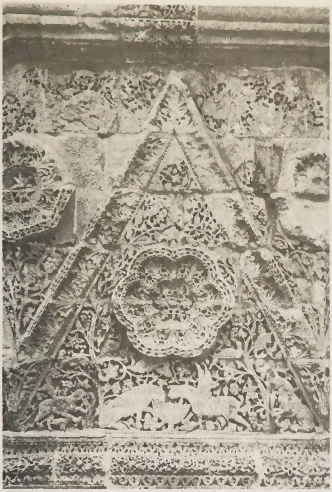
The themes and motifs of early Islāmic decoration can be divided into three major groups. The first kind of ornamentation simply emphasizes the shape or contour of an architectural unit. The themes used were vegetal bands for vertical or horizontal elements, marble imitations for the lower parts of long walls, chevrons or other types of borders on floors and domes, and even whole trees on the spandrels or soffits (undersides) of arches as in the Umayyad Mosque of Damascus or the Dome of the Rock; all these motifs tend to be quite traditional, being taken from the rich decorative vocabularies of pre-Islāmic Iran or of the ancient Mediterranean world.

Three kinds of early architectural decoration

The second group consists of decorative motifs for which a concrete iconographic meaning can be given. In the Dome of the Rock and the Umayyad Mosque of Damascus, as well as possibly the mosques of Córdoba and of Medina, there were probably iconographic programs. It has been shown, for example, that the huge architectural and vegetal decorative motifs at Damascus were meant to symbolize a sort of idealized paradise on earth, while the crowns of the Jerusalem sanctuary are thought to have been symbols of empires conquered by Islām. But it is equally certain that this use of visual forms in mosques for ideological and symbolic purposes was not easily accepted, and most later mosques are devoid of iconographically significant themes. The only exceptions fully visible are the Qurʾānic inscriptions in the mosque of Ibn Ṭūlūn at Cairo, which were used both as a reminder of the faith and as an ornamental device to emphasize the structural lines of the building. Thus the early Islāmic mosque eventually became austere in its use of symbolic ornamentation, with the exception of the *miḥrāb,* which was considered as a symbol of the unity of all believers.

Like religious architecture, secular buildings seem to have been less richly decorated at the end of the early Islāmic period than at the beginning. The paintings, sculptures, and mosaics of Qaṣr al-Ḥayr West, Khirbat al-Mafjar, Qaṣr ʿAmrah, and Sāmarrāʾ primarily illustrated the life of the prince. There were official iconographic compositions, such as the monarch enthroned, or ones of pleasure and luxury, such as hunting scenes or depictions of the prince surrounded by dancers, musicians, acrobats, and unclad women. Few of these so-called princely themes were iconographic inventions of the Muslims. They usually can be traced back either to the classical world of ancient Greece and Rome or to pre-Islāmic Iran and Central Asia.

The third type of architectural decoration consists of large panels, most often in stucco, for which no meaning or



Triangle stone relief from the facade of Mshattā in Transjordan, early 8th century. In the Islamisches Museum, Staatliche Museen zu Berlin.
By courtesy of the Islamisches Museum, Staatliche Museen zu Berlin

interpretation is yet known. These panels might be called ornamental in the sense that their only apparent purpose was to beautify the buildings in which they were installed, and their relationship to the architecture is arbitrary. The Mshattā facade's decoration of a huge band of triangles is, for instance, quite independent of the building's architectural parts. Next to Mshattā, the most important series of examples of the third type of ornamentation come from Sāmarrā', although striking examples are also to be found at Khirbat al-Mafjar, Qaṣr al-Ḥayr East and West, al-Fusṭāṭ, Sīrāf, and Nīshāpūr. Two decorative motifs were predominately used on these panels: a great variety of vegetal motifs and geometric forms. At Sāmarrā' these panels eventually became so abstract that individual parts could no longer be distinguished, and the decorative design had to be viewed in terms of the relationships between line and shape, light and shade, horizontal and vertical axes, and so forth. Copied consistently from Morocco to Central Asia, the aesthetic principles of this latter type of a complex overall design influenced the development of the principle of arabesque ornamentation.

Islāmic architectural ornamentation does not lend itself easily to chronological stylistic definition. In other words, it does not seem to share consistently a cluster of formal characteristics. The reason is that in the earliest Islāmic buildings the decorative motifs were borrowed from an extraordinary variety of stylistic sources: classical themes illusionistically rendered (*e.g.*, the mosaics of the Umayyad Mosque of Damascus), hieratic Byzantine themes (*e.g.*, the Umayyad Mosque of Damascus and Qaṣr 'Amrah), Sāsānian motifs, Central Asian motifs (especially the sculpture from Umayyad palaces), and the many regional styles of ornamentation that had developed in all parts of the pre-Islāmic world. It is the wealth of themes and motifs, therefore, that constitutes the Umayyad style of architectural decoration. The 'Abbāsids, on the other hand, began to be more selective in their choice of ornamentation.

Decorative arts. Very little is known about early Islāmic gold and silver objects, although their existence is mentioned in many texts as well as suggested by the wealth of the Muslim princes. Except for a large number of silver plates and ewers belonging to the Sāsānian tradition, nothing has remained. These silver objects were probably made for Umayyad and 'Abbāsid princes, although there is much controversy among scholars regarding their authenticity and date of manufacture.

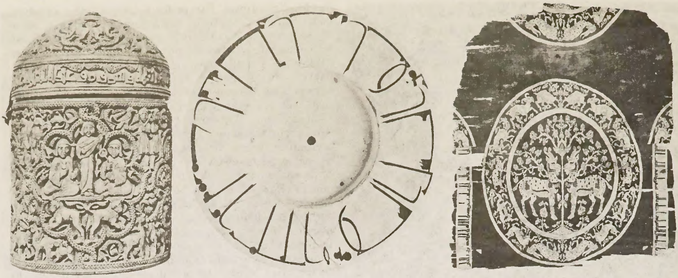
For entirely different reasons it is impossible to present any significant generalities about the art of textiles in the early Islāmic period. Problems of authenticity are few. Dating from the 10th century are a large number of Būyid silks, Buyid silks
a group of funerary textiles with plant and animal motifs as well as poetic texts. Very little order has yet been made of an enormous mass of often well-dated textile fragments, and therefore, except for the Būyid silks, it is still impossible to identify any one of the textile types mentioned in early medieval literary sources. Furthermore, since it can be assumed that pre-Islāmic textile factories were taken over by the Muslims and since it is otherwise known that textiles were easily transported from one area of the Muslim world to the other or even beyond it, it is still very difficult to define Islāmic styles as opposed to Byzantine or to Coptic ones. The obvious exception lies in those fragments that are provided with inscriptions, and the main point to make is therefore that one of the characteristic features of early Islāmic textiles is their use of writing for identifying and decorative purposes. But, while true, this point in no way makes it possible to deny an Islāmic origin to fragments that are not provided with inscriptions, and thus one must await further investigations of detail before being able to define early Islāmic textiles.

The most important medium of early Islāmic decorative Early ceramics
arts is pottery. Initially Muslims continued to sponsor whatever varieties of ceramics had existed before their arrival. Probably in the last quarter of the 8th century new and more elaborate types of glazed pottery were produced. This new development did not replace the older and simpler types of pottery but added a new dimension to the art of Islāmic ceramics. Because of the still incompletely published studies on the unfinished excavations carried out at Nīshāpūr, Sīrāf, Qaṣr al-Ḥayr East, and al-Fusṭāṭ, the scholarship on these ceramics is likely to be very much modified. Therefore, this section will treat only the most general characteristics of Islāmic ceramics, avoiding in particular the complex archaeological problems posed by the growth and spread of individual techniques.

The area of initial technical innovation seems to have been Iraq. Trade with Central Asia brought Chinese ceramics to Mesopotamia, and Islāmic ceramicists sought to imitate them. It is probably in Iraq, therefore, that the technique of lustre glazing was first developed in the Muslim world. This gave the surface of a clay object a metallic, shiny appearance. Egypt also played a leading part in the creation of the new ceramics. Since the earliest datable lustre object (a glass goblet with the name of the governor who ruled in 773, now in the Cairo Museum of Islamic Art) was Egyptian, some scholars feel that it was in Egypt and not Iraq that lustre was first used. Early pottery was also produced in northeastern Iran, where excavations at Afrāsiyāb (Samarkand) and Nīshāpūr have brought to light a new art of painted underglaze pottery. Its novelty was not so much in the technique of painting designs on the slip and covering them with a transparent glaze as in the variety of subjects employed.

While new ceramic techniques may have been sought to imitate other mediums (mostly metal) or other styles of pottery (mostly Chinese), the decorative devices rapidly became purely and unmistakably Islāmic in style. A wide variety of motifs were combined: vegetal arabesques or single flowers and trees; inscriptions, usually legible and consisting of proverbs or of good wishes; animals that were usually birds drawn from the vast folkloric past of the Near East; occasionally human figures drawn in a strikingly abstract fashion; geometric designs; all-over abstract patterns; single motifs on empty fields; and simple splashes of colour, with or without underglaze sgraffito designs (i.e., designs incised or sketched on the body or the slip of the object). All of these motifs were used on both the high-quality ceramics of Nīshāpūr and Samarkand as well as on Islāmic folk pottery.

Although ceramics has appeared to be the most character-



Early Islāmic decorative arts.
(Left) Ivory casket made for al-Mughīrah, son of ʿAbd ar-Raḥmān III, from Córdoba, Spain, 968. In the Louvre, Paris. Ht. 15 cm. (Centre) Bowl of Samarkand ware with calligraphic decoration, 10th century. In the Louvre, Paris. Diameter 37.5 cm. (Right) Fragment of a silk tomb cover with a woven design of pairs of ibex, from Iran, 998. In the Cleveland Museum of Art, Ohio. 78.1 cm × 66 cm.
By courtesy of (left, centre) the Musee du Louvre, Paris, (right) the Cleveland Museum of Art, Purchase from the J.H. Wade Fund; photographs, (left) Mansell—Giraudon from Art Resource, (centre) Cliche Musees Nationaux, Paris

istic medium of expression in the decorative arts during the early Islāmic period, it has only been because of the greater number of preserved objects. Glass was as important, but examples have been less well preserved. A tradition of ivory carving developed in Spain, and the objects dating from the last third of the 10th century onward attest to the high quality of this uniquely Iberian art. Many of these carved ivories certainly were made for princes; therefore it is not surprising that their decorative themes were drawn from the whole vocabulary of princely art known through Umayyad painting and sculpture of the early 8th century. These ivory carvings are also important in that they exemplify the fact that an art of sculpture in the round never totally disappeared in the Muslim world—at least in small objects.

Spanish ivory carving

Assessment. There are three general points that seem to characterize the art of the early Islāmic period. It can first be said that it was an art that sought self-consciously, like the culture sponsoring it, to create artistic forms that would be identifiable as being different from those produced in preceding or contemporary non-Islāmic artistic traditions. At times, as in the use of the Greco-Roman technique of mosaics or in the adoption of Persian and Roman architectural building technology, early Islāmic art simply took over whatever traditions were available. At other times, as in the development of the mosque as a building type, it recomposed into new shapes the forms that had existed before. On the other hand, in ceramics or the use of calligraphic ornamentation, the early Islāmic artist invented new techniques and a new decorative vocabulary. Whatever the nature of the phenomenon, it was almost always an attempt to identify itself visually as unique and different. Since there was initially no concept about what should constitute an Islāmic tradition in the visual arts, the early art of the Muslims often looks like only a continuation of earlier artistic styles, forms, subjects, and techniques. Many mosaics, silver plates, or textiles, therefore, were not considered to be Islāmic until recently. In order to be understood, then, as examples of the art of a new culture, these early buildings and objects have to be seen in the complete context in which they were created. When so seen they appear as conscious choices by the new Islāmic culture from its immense artistic inheritance.

Question of an early Islāmic style

A second point of definition concerns the question of whether there is an early Islāmic style or perhaps even several styles in some sort of succession. The fascinating fact is that there is a clear succession only in those artistic features that are Islāmic inventions—nonfigurative ornament and ceramics. For it is only in development of these features that one can assume to find the conscious search for form that can create a period style. Elsewhere, especially in palace art, the Muslim world sought to relate itself to an earlier and more universal tradition of princely art; its monuments, therefore, are less Islāmic than typological. In the new art of the Muslim bourgeoisie, however, uniquely Islāmic artistic phenomena began to evolve.

Finally, the geographical peculiarities of early Islāmic art must be reiterated. Its centres were Syria, Iraq, Egypt, northwestern Iran, and Spain. Of these, Iraq was probably the most originally creative, and it is from Iraq that a peculiarly Islāmic visual koine (a commonly accepted and understood system of forms) was derived and spread throughout the Islāmic world. This development, of course, is logical since the capital of the early empire and some of the first purely Muslim cities were in Iraq. In western Iran, in Afghanistan, in northern Mesopotamia, and in Morocco the more atypical and local artistic traditions were more or less affected by the centralized imperial system of Iraq. This tension between a general pan-Islāmic vocabulary and a variable number of local vocabularies was to remain a constant throughout the history of Islāmic art and is certainly one of the reasons for the difficulty, if not impossibility, one faces in trying to define an Islāmic style.

MIDDLE PERIOD

The middle period in the development of Islāmic art extends roughly from the year 1000 to 1500, when a strong central power with occasional regional political independence was replaced by a bewildering mosaic of overlapping dynasties. Ethnically this was the time of major Turkish and Mongol invasions that brought into the Muslim world new peoples and institutions. At the same time, Berbers, Kurds, and Iranians, who had been within the empire from the beginning of Islām, began to play far more effective historical and cultural roles, shortlived for the Kurds, but uniquely important for the Iranians. Besides political and ethnic confusion, there was also religious and cultural confusion during the middle period. The 10th century, for example, witnessed the transformation of the Shīʿite heterodoxy into a major political and possibly cultural phenomenon, while the extraordinary development taken by the personal and social mysticism known as Ṣūfism modified enormously the nature of Muslim piety. Culturally the most significant development was perhaps that of Persian literature as a highly original new verbal expression existing alongside the older Arabic literary tradition. Finally, the middle period was an era of expansion in all areas except Spain, which was completely lost to the Muslims in 1492 with the conquest of the Kingdom of Granada by Ferdinand II and Isabella. Anatolia and the Balkans, the Crimea, much of Central Asia and northern India, and parts of eastern Africa all became new Islāmic provinces. In some cases this expansion was the result of conquests, but in others it had been achieved through missionary work.

The immense variety of impulses that affected the Muslim world during these five centuries was one of the causes of the bewildering artistic explosion that also characterizes the middle period. Although much work has been done on individual monuments, scholarship is still in its infancy. It is particularly difficult, therefore, to decide on the appropriate means of organizing this information: by geographical or cultural areas (*e.g.*, Iran, Egypt, Morocco), by individual dynasties (*e.g.*, Seljuqs, Timurids), by periods (*e.g.*, 13th century before the Mongol invasions), or even by social categories (*e.g.*, the art of princes, the art of cities). Thus, the five following divisions of Fāṭimid, Seljuq, Western Islāmic, Mamlūk, and Mongol Iran (Il-Khanid and Timurid) art are partly arbitrary and to a large extent tentative. Their respective importance also varies, for what is known as Seljuq art certainly overwhelms almost all others in its importance.

Fāṭimid art (909–1171). The Fāṭimids were technically an Arab dynasty professing with missionary zeal the beliefs of the Ismāʿīlī sect of the Shīʿite branch of Islām. The dynasty was established in Tunisia and Sicily in 909. In 969 the Fāṭimids moved to Egypt and founded the city of Cairo. They soon controlled Syria and Palestine. In the latter part of the 11th century, however, the Fāṭimid empire began to distintegrate internally and externally; the final demise occurred in 1171. But it is not known which of the obvious components of the Fāṭimid world was more significant in influencing the development of the visual arts: its heterodoxy, its Egyptian location, its missionary relationship with almost all provinces of Islām, or the fact that during its heyday in the 11th century it was the only wealthy Islāmic centre and could thus easily gather artisans and art objects from all over the world.

Architecture. The great Fāṭimid mosques of Cairo—al-Azhar (started in 970) and al-Ḥākim (*c.* 1002–03)—were designed in the traditional hypostyle plan with axial cupolas. It is only in such architectural details as the elaborately composed facade of al-Ḥākim, with its corner towers and vaulted portal, that innovations appear, for most earlier mosques did not have large formal gates, nor was much attention previously given to the composition of the exterior facade. The Fāṭimids' architectural traditionalism was certainly a conscious attempt to perpetuate the existing aesthetic system.

Architectural traditionalism

Although much less is known about it, the Great Palace of the Fāṭimids belonged to the tradition of the enormous palace-cities typical of the ʿAbbāsids. Mediterranean rather than Iranian influences, however, played a greater part in the determination of its uses and functions. The whole city of Cairo (Arabic: al-Qāhirah, meaning "the Victorious"), on the other hand, has many symbolic and visual aspects that suggest a willful relationship to Baghdad.

The originality of Fāṭimid architecture does not lie in works sponsored by the caliphs themselves, even though Cairo's well-preserved gates and walls of the second half of the 11th century are among the best examples of early medieval military architecture. It is rather the patronage of lower officials and of the bourgeoisie, if not even of the humbler classes, that was responsible for the most interesting Fāṭimid buildings. The mosques of al-Aqmar (1125) and of aṣ-Ṣāliḥ (*c.* 1160) are among the first examples of monumental small mosques constructed to serve local needs. Even though their internal arrangement is quite traditional, their plans were adapted to the space available in the urban centre. These mosques were elaborately decorated on the exterior, exhibiting a conspicuousness absent from large hypostyle mosques.

A second innovation in Fāṭimid architecture was the tremendous development of mausoleums. This may be explained partially by Shīʿism's emphasis on the succession of holy men, but the development of these buildings in terms of both quality and quantity indicates that other influential social and religious issues were also involved. Most of the mausoleums were simple square buildings surmounted by a dome. Many of these have survived in Cairo and Aswān. Only a few, such as the *mashhad* at Aswān, are somewhat more elaborate, with side rooms. The most original of these commemorative buildings is the Juyūshī Mosque (1085) overlooking the city of Cairo. Properly speaking, it is not a mausoleum but a monument celebrating the reestablishment of Fāṭimid order after a series of popular revolts.

The Fāṭimids introduced, or developed, only two major constructional techniques: the systematization of the four-centred "keel" arch and the squinch. The latter innovation is of greater consequence because the squinch became the most common means of passing from a square to a dome, although pendentives were known as well. A peculiarly Egyptian development was the *muqarnas* squinch, which consisted of four units: a niche bracketed by two niche segments, superimposed with an additional niche. The complex profile of the *muqarnas* became an architectural element in itself used for windows, while the device of using niches and niche segments remained typical of Egyptian decorative design for centuries. It still is impossible to say whether the *muqarnas* was invented in Egypt or inspired by other architectural traditions (most likely Iranian). Fāṭimid domes were smooth or ribbed and developed a characteristic "keel" profile.

Use of the squinch

In the use of materials (brick, stone, wood) and structural concepts, Fāṭimid architecture continued earlier traditions. Occasionally local styles were incorporated, among them features of Tunisian architecture in the 10th century or of upper Mesopotamian in the late 11th century.

Stone sculpture, stucco work, and carved wood were utilized for architectural decorations. The Fāṭimids also employed mosaicists, who mostly worked in places like Jerusalem, where they imitated or repaired earlier mosaic murals. Many fragments of Fāṭimid wall paintings have survived in Egypt. Most of them, however, are too small to allow for making any iconographic or stylistic conclusions, with the exception of the mid-12th-century ceiling of the Cappella Palatina at Palermo. Built by the Norman kings of Sicily, the palace chapel was almost certainly decorated by Fāṭimid artists, or at least the artists adhered to Fāṭimid models. The hundreds of facets in the *muqarnas* ceiling were painted, notably with many purely ornamental vegetal and zoomorphic designs but also with scenes of daily life and many subjects that have not yet been explained. Stylistically influenced by Iraqi ʿAbbāsid art, these paintings are innovative in their more spatially aware representation of personages and of animals. Very similar tendencies appear also in the stucco and wood sculptures of Fāṭimid decoration. The stunning abstraction of the architectural decoration at Sāmarrāʾ tends to give way to more naturalistically conceived vegetal and animal designs; occasionally whole narrative scenes appear carved on wood. Another decorative trend is especially used on 12th-century *miḥrāb*s: explicitly complicated geometric patterns, usually based on stars, which in turn generate octagons, hexagons, triangles, and rectangles. Geometry becomes a sort of network in the midst of which small vegetal units continue to remain, often as inlaid pieces. Long inscriptions written in very elaborate calligraphies also became a typical form of architectural decoration on most of the major Fāṭimid buildings.

A clear separation must be made between the decorative arts sought by Fāṭimid princes and the arts produced within their empire. Little has been preserved of the former, notably a small number of superb ewers in rock crystal. A text has survived, however, that describes the imperial treasures looted in the middle of the 11th century by dissatisfied mercenary troops. It lists gold, silver, enamel, and porcelain objects that have all been lost, as well as textiles (perhaps the cape of the Norman king Roger II [Kunsthistorisches Museum, Vienna] is an example of the kind of textiles found in this treasure). The inventory also records that the Fāṭimids had in their possession many works of Byzantine, Chinese, and even Greco-Roman provenance. Altogether, then, it seems that the imperial art of the Fāṭimids was part of a sort of international royal taste that downplayed cultural or political differences.

Ceramics, on the other hand, were primarily produced by local urban schools and were not an imperial art. The most celebrated type of Fāṭimid wares were lustre-painted ceramics from Egypt itself. A large number of artisans' names have been preserved, thereby indicating the growing prestige of these craftsmen and the aesthetic importance

Lustreware with figural decoration

of their pottery. Most of the surviving lustre ceramics are plates on which the decoration of the main surface has been emphasized. The decorative themes used were quite varied and included all the traditional Islāmic ones: *e.g.*, calligraphy, vegetal and animal motifs, arabesques. The most distinguishing feature of these Fāṭimid ceramics, however, is the representation of the human figure. Some of these ceramics have been decorated with sim-

Fāṭimid lustre-painted dish depicting a cockfight, from Egypt, late 11th–early 12th century. In the Edmund de Unger Collection, London. Diameter 24.1 cm.

By courtesy of the Edmund de Unger Collection; photograph, A.C. Cooper Ltd.

plified copies of illustrations of the princely themes, but others have depictions of scenes of Egyptian daily life. The style in which these themes have been represented is simultaneously the hieratic, ornamental manner traditional to Islāmic painting combined with what can almost be called spatial illusionism. Wheel-cut rock crystal, glass, and bronze objects, especially animal-shaped aquamaniles (a type of water vessel) and ewers, are also attributed to the Fāṭimids.

Book illustration. Manifestations of nonprincely Fāṭimid art also included the art of book illustration. The few remaining fragments illustrate that probably after the middle of the 11th century there developed an art of representation other than the style used to illustrate princely themes. This was a more illusionistic style that still accompanied the traditional ornamental one in the same manner as in the paintings on ceramics.

Transitional role of Fāṭimid art

In summary it would appear that Fāṭimid art was a curiously transitional one. Although much influenced by earlier Islāmic and non-Islāmic Mediterranean styles, the Fāṭimids devised new structural systems and developed a new manner of painting representational subjects, which became characteristic of all Muslim art during the 12th century. Neither documentary nor theoretical research in Islāmic art, however, has developed sufficiently to clearly establish whether the Fāṭimids were indeed innovators or whether their art was a local phenomenon that is only accidentally relatable to what followed.

Seljuq art. During the last decades of the 10th century, at the Central Asian frontiers of Islām, a migratory movement of Turkic peoples began that was to affect the whole Muslim world up to and including Egypt. The dominant political force among these Turks was the dynasty of the Seljuqs, but it was not the only one; nor can it be demonstrated, as far as the arts are concerned, that it was the major source of patronage in the period to be discussed anywhere but in Anatolia in the 12th and 13th centuries. The Seljuq empire, therefore, consisted of a succession of dynasties, and all but one (the Ayyūbids of Syria, Egypt, and northern Mesopotamia) were Turkic.

A complex feudal system was established and centred on urban areas. Cities were established or expanded, particularly in western Iran, Anatolia, and Syria. Militant Muslims, the Seljuqs also sought to revive Muslim orthodoxy. Although politically unruly and complicated in their relationships to one another, the successive and partly overlapping dynasties of the Ghaznavids, Ghūrids, the Great Seljuqs, Qarakhānids, Zangids, Ayyūbids, Seljuqs of Rūm, and Khwārezm-Shāhs (considering only the major ones) seem to have created a comparatively unified culture from India to Egypt. The art of the Seljuq period, however, is difficult to discuss coherently both because of the wealth of examples and because of the lack of synchronization between various technical and regional developments. This complex world fell apart under the impact of the Mongol invasions that, from 1220 until 1260, swept through the Muslim lands of the Near East.

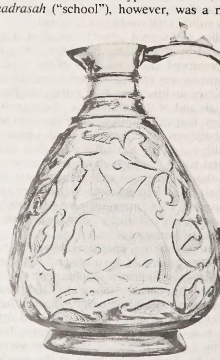
Architecture. The functions of monumental architecture in the Seljuq period were considerably modified. Large congregational mosques were still built. The earliest Seljuq examples occur in the two major new provinces of Islām—Anatolia and northwestern India—as well as in the established Muslim region of western Iran. In some areas, such as the Isfahan region, congregational mosques were rebuilt, while in other parts of Islām, such as Syria or Egypt, where there was no need for new large mosques, older ones were repaired and small ones were built. The latter were partly restricted to certain quarters or groups or were commissioned by various guilds, particularly in Damascus.

A curious side aspect of the program of building, rebuilding, or decorating mosques was the extraordinary development of minarets. Particularly in Iran, dozens of minarets are preserved from the 12th and 13th centuries, while the mosques to which they had been attached have disappeared. It is as though the visual function of the minaret was more important than the religious institution to which it was attached.

Small or large, mausoleums increased in numbers and became at this time the ubiquitous monument they appear to be. Most of the mausoleums, such as the tomb tower of Abū Yazīd al-Bisṭāmī (died 874) at Besṭām, were dedicated to holy men—both contemporary Muslim saints and all sorts of holy men dead for centuries (even pre-Islāmic holy men, especially biblical prophets, acquired a monument). The most impressive mausoleums, however, ones like the one of Sanjar at Merv, were built for royalty. Pilgrimages were organized and in many places hardly mentioned until then as holy places (*e.g.*, Meshed, Besṭām, Mosul, Aleppo); a whole monastic establishment serving as a centre for the distribution of alms was erected with hostels and kitchens for the pilgrims.

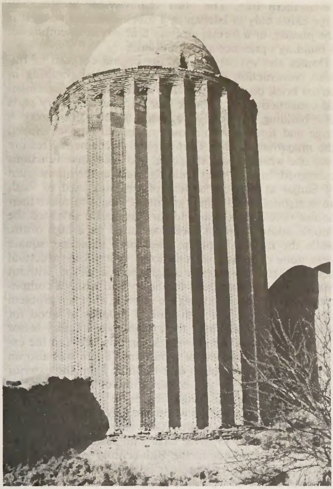
Origin and development of the *madrasah*

Although enormously expanded, mosques, minarets, and mausoleums were not new types of Islāmic architecture. The *madrasah* ("school"), however, was a new building



Wheel-cut rock crystal ewer from Egypt, 11th century. In the Victoria and Albert Museum, London. Ht. 21.5 cm.

By courtesy of the Victoria and Albert Museum, London



Tomb tower at the shrine of Abū Yazīd al-Bisṭāmī at Besṭām, Iran, 1313.
Josephine Powell, Rome

type. There is much controversy as to why and how it really developed. Although early examples have been discovered in Iran, such as the 11th-century *madrasah* of Khargird in Iran and at Samarkand, it is from Anatolia, Syria, and Egypt that most of the information about the *madrasah* has been derived. In the latter regions it was usually a privately endowed establishment reserved for one or two of the schools of jurisprudence of orthodox Islām. It had to have rooms for teaching and living quarters for the students and professors. Often the tomb of the founder was attached to the *madrasah*. Later *madrasahs* were built for two or three schools of jurisprudence, and the Mustanṣiriyah in Baghdad was erected in 1233 to be a sort of ecumenical *madrasah* for the whole of Sunnī Islām.

In the Seljuq period there occurred a revival of the *ribāṭ* inside cities. *Khānqāhs*, monasteries, and various establishments of learning other than formal *madrasahs* were also built.

An impressive development of secular architecture occurred under the Seljuqs. The most characteristic building of the time was the citadel, or urban fortress, through which the new princes controlled the usually alien city they held in fief. The largest citadels, like those of Cairo and Aleppo, were whole cities with palaces, mosques, sanctuaries, and baths. Others, like the Citadel of Damascus, were simpler constructions. Occasionally, as in the Euphrates valley, single castles were built, possibly in imitation of those constructed by the Christian crusaders. Walls surrounded most cities, and all of them were built or rebuilt during the Seljuq period.

Little is known about Seljuq palaces or private residences in general. A few fragments in Konya or in Mosul are insufficient to give a coherent idea about urban palaces, and it is only in Anatolia and in Central Asia that an adequate idea of other types can be obtained. Anatolian palaces are on the whole rather small villa-like establishments; but, in Afghanistan and Central Asia, excavations at Tirmidh, Lashkarī Bāzār, and Ghaznī have brought to light a whole group of large royal palaces erected in the 11th and early 12th centuries.

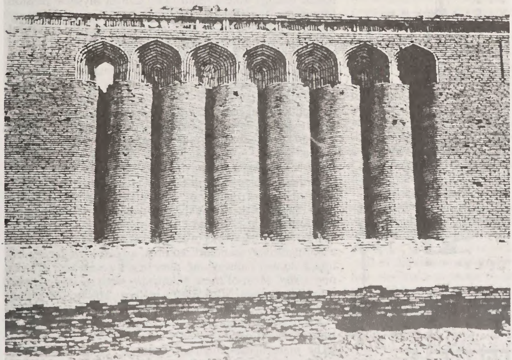
Seljuq commercial architecture

Commercial architecture became very important. Individual princes and cities probably were trying to attract business by erecting elaborate caravansaries on the main trade routes such as Rebāṭ-e Malek built between Samarkand and Bukhara in Iran. The most spectacular caravansaries were built in the 13th century in Anatolia. Equally impressive, however, although less numerous, are the caravansaries erected in eastern Iran and northern Iraq. Bridges also were rebuilt and decorated like the one at Cizre in Turkey.

The forms of architecture developed by the Seljuqs were remarkably numerous and varied considerably from region to region. Since the Iranian innovations dating from the 11th century and first half of the 12th century are the earliest and, therefore, probably influenced all other areas of the Seljuq empire, they will be discussed first.

Even though it is not entirely typical, the justly celebrated Great Mosque of Isfahan was one of the most influential of all early Seljuq religious structures. Probably completed around 1130 after a long and complicated

By courtesy of the General Direction of Museums and Historical Monuments, Ministry of Culture and Arts, Tehran, Iran



Brickwork facade of the caravansary of Rebāṭ-e Malek, 11th century.

Use of the *eyvān*

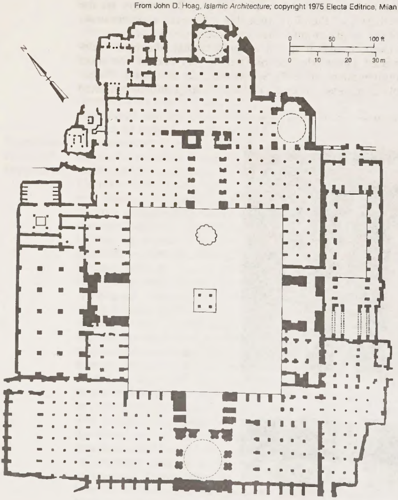
history of rebuildings, it consisted of a large courtyard on which opened four large vaulted halls known as *eyvāns*; the *eyvāns* created the compositional axes of each side of the court. On the side of the *qiblah* the hall of the main *eyvān* was followed by a huge cupola. The area between *eyvāns* was subdivided into a large number of square bays covered by domes. The Isfahan mosque also had a unique feature: on the north side a single domed hall positioned on the main axis of the building was in all probability a formal hall for princes to change their clothes before entering into the sanctuary of the mosque.

The two features of the Great Mosque at Isfahan that became characteristic of Seljuq mosques were the *eyvān* and the dome. The *eyvān* was an architectural element known already in Sāsānian architecture that had been used in residential buildings from Egypt to Central Asia before the 11th century. In fact, the use of the *eyvān* was not restricted to just mosques, but it also appears in palaces (Lashkarī Bāzār), caravansaries (Rebāṭ-e Sharaf), and in *madrasahs*. The *eyvān* was, in other words, a unit of architectural composition that had no specific use and, therefore, no meaning. In the mosques of the 12th century, four *eyvāns* were used, at least in the clearly definable architectural school of western Iran (*e.g.*, Ardestān, Zavāreh). This kind of composition had two principal effects. One was that the *eyvāns* centralized the visual effect of the mosque by making the courtyard the centre of the building. The other effect of this composition was that it broke up into four areas what had for centuries been a characteristic of the mosque: its single, unified space. The reasons for these developments are still speculative.

Whether large or small, cupolas or domes were used in mosques, caravansaries, and palaces. They were the main architectural features of almost all mausoleums, where they were set over circular or polygonal rooms.

Two characteristic Iranian architectural forms are not present in the Great Mosque of Isfahan but occur elsewhere in the city. One is the tower. Those narrow and tall (up to about 150 feet [50 metres]) were minarets, of which several dozen have been preserved all over Iran and Central Asia (such as the one at Jām). Shorter and squatter towers were mausoleums. These were particularly typical

From John D. Hoag, *Islamic Architecture;* copyright 1975 Electa Editrice, Milan



Present-day plan of the Great Mosque at Isfahan.

of northern Iran. The other characteristic architectural type exists only in Isfahan in a much-damaged state. It is the *pīshṭāq*, or a formal gateway that served to emphasize a building's presence and importance.

The *pīshṭāqs* of Isfahan

Domes and *eyvāns* indicate the central concern of Iranian construction during the Seljuq period: vaulting in baked brick became the main vehicle for any monumental construction (mud brick was used for secondary parts of a building, frequently for certain secular structures). A large and forcefully composed octagonal base developed the *muqarnas* squinch from a purely ornamental feature into one wherein both structural and decorative functions combined. In some later buildings, such as the mausoleum of Sanjar at Merv, a system of ribs was used to vault an octagonal zone. Seljuq architects sought to make their domes visible from afar and for this reason invented the double dome. Its outer shell was raised on a high drum, while the interior kept the traditional sequence: square base, zone of transition, and dome. Using this structural device, therefore, exterior height was achieved without making the exterior dome too heavy and without complicating the task of decorating the interior, always a problem in countries like Iran with limited supplies of wood for scaffolding. Domes along the *eyvāns* were another factor in contributing to the growing separation between the exterior and interior view of a building. There was also an emphasis on the visibility of a building from the exterior that is indicated by the construction of tall circular or polygonal minarets and high facades.

Architectural decoration was intimately tied to structure. Two mediums predominated. One was stucco, which continued to be used to cover large wall surfaces. The other was brick. Originating in the 10th-century architecture of northeastern Iran, brick came to be employed as a medium of construction as well as a medium of decoration. The complex decorative designs worked out in brick often had a rigidly geometric effect. Especially cut shapes of terracotta and brick, frequently produced in unusual sizes, served to soften these geometric patterns by modifying their tactile impact and by introducing additional curved or beveled lines to the straight lines of geometry.

Paintings were used for architectural decoration, especially in palaces. From the second half of the 12th century coloured tiles began to be utilized to emphasize the contour of a decorative area in a structural unit; tiles were not used, however, to cover whole walls. There are also examples of architectural sculpture of animals and people.

Most of the decorative designs tended to be subordinated to geometry, and even calligraphic or vegetal patterns were affected by a seemingly mathematically controlled aesthetic. It has been suggested that these complex geometric designs were a result of an almost mystical passion for number theories that were popularized in 11th-century Iran by such persons as the scholar and scientist al-Bīrūnī or the poet-mathematician Omar Khayyam. But even if the impulses for geometric design were originally created at the highest intellectual level, the designs themselves rapidly became automatic patterns. Their quality was generally high, but a tendency toward facility can be observed in such buildings as Rebāṭ-e Sharaf.

In Iraq, northern Mesopotamia, Syria, and Egypt (after 1171), the architectural monuments do not, on the whole, appear as overwhelmingly impressive as those of Iran, largely because the taste of Umayyad and ʿAbbāsid times continued to dominate mosque architecture. It is in the construction of new building types, particularly the *madrasah*, that the most originality is apparent. The Syrian *madrasahs* in Damascus, like al-ʿĀdilīyah, aẓ-Ẓāhirīyah, or the works of Nureddin, tended also to follow a comparatively standardized plan: an elaborate facade led into a domed hallway and then into a court with at least one *eyvān*. Most of these *madrasahs* were small and were fitted into a preexisting urban pattern. The use of *eyvāns* and the construction of the many minarets found in Mosul or on the Euphrates certainly attest to the influence of Iranian Seljuq design.

The main achievement of Ayyūbid, Zangid, or Seljuq architecture in the Fertile Crescent was the translating into stone of new structural systems first developed in brick. The most impressive instance of this lies in the technically complex *muqarnas* domes and half domes or in the *muqarnas* pendentives of Syrian buildings. Elaborate *miḥrāb*s were also made of multicoloured stones that were carefully cut to create impressive patterns. The architecture of the Fertile Crescent, therefore, was still dominated by the sheer force of stone as a material for both construction as well as decoration, and, therefore, the architecture was more Mediterranean in effect than were the buildings of Iran.

The hybrid style of Anatolia

This Mediterranean tendency was also evident in the 13th-century architecture of Seljuq Anatolia. This new province of Islām was rapidly populated with new immigrants and consequently gathered themes and motifs from throughout the Muslim world, as well as from the several native Anatolian traditions of Byzantine, Armenian, and Georgian architecture. The resulting assimilation of styles produced an overwhelmingly original architecture, for each building in Konya, Kayseri, Sivas, Divriği, Erzurum, or on the roads between them is a unique monument.

Functionally the buildings in Anatolia do not differ from those in other parts of the Muslim world. All the structural forms found in Syria and Iran can be found in Anatolia as well, although they have often been adapted to local materials. Three uniquely Anatolian architectural features, however, can be distinguished. One was limited to Konya at this time but would have an important widespread development later on. As it appears in the Ince or Karatay *medreses*, it consists of the transformation of the central courtyard into a domed space while maintaining the *eyvān*. Thus the centralized aspect of the *eyvān* plan becomes architecturally explicit. The second feature is the creation of a facade that usually consisted of a high central portal—often framed by two minarets—with an elaborately sculpted decorative composition that extended to two corner towers. The third distinguishing feature of Anatolian Seljuq architecture is the complexity of the types of funerary monuments that were constructed.

From the point of view of construction, most of Anatolian architecture is of stone. In Konya and a number of eastern Anatolian instances, brick was used. Barrel vaults, groin vaults, *muqarnas* vaults, squinch domes, pendentive domes, and the new pendentive known as "Turkish triangle" (a transformation of the curved space of the traditional pendentive into a fanlike set of long and narrow triangles built at an angle from each other) were all used by Anatolian builders, thereby initiating the great development of vault construction in Ottoman architecture (see below).

Ara Guler

Ince Minare at Konya, Tur., 1258. Detail view showing the sculptural ornamentation of the main facade portal and the decorative brickwork of the minaret.

Architectural decoration

Architectural decoration consisted primarily in the stone sculpture found on the facades of religious and secular buildings. Although influenced by Iran and Syria in many details, most Anatolian themes were original, although some exhibit Armenian and possibly Western influences. The exuberance of Anatolian architectural decoration can perhaps be best demonstrated in the facades of Sivas' Gök Medrese and of Konya's Ince Minare. In addition to the traditional geometric, epigraphic, and vegetal motifs, a decorative sculpture in the round or in high relief was created that included many representations of human figures and especially animals. Whether this sculpture is essentially a reflection of the decorative wealth of pre-Islāmic monuments in Anatolia, or whether it is the vestige of a pagan Turkish art that originated in Central Asia, is still an unsolved historical problem.

There are few examples of wall painting from Anatolia. Especially in Konya, however, a major art of painted-tile decoration did evolve, possibly developed by Iranian artists who fled from the Mongol onslaught.

In summing up the architectural development of the Seljuq period, three points seem to be particularly significant. One is the expansion of building typology and the erection of new monumental architectural forms, thus illustrating an expansion of patronage and a growing complexity of taste. The second point is that, regardless of the quality and interest of monuments in the Fertile Crescent, Egypt, and Anatolia, the most inventive and exciting architecture in the 11th and 12th centuries was that of Iran. But, far more than in the preceding period, regional needs and regional characteristics seem to predominate over synchronic and pan-Islāmic ones. Finally, there was a striking growth of architectural decoration both in sophistication of design and in variation of technique.

Other arts. Although probably not as varied as architecture, the other arts of the Seljuq period also underwent tremendous changes. They demonstrate an extraordinary artistic energy, a widening of the social patronage of the arts, and a hitherto unknown variety of topics and modes of expression. It was as though the Seljuq period was gathering a sort of aesthetic momentum, but this effort seems to have been curtailed by the Mongol invasion. Chronologically, almost all surviving documentation and examples of these arts date from the latter part of the period, after 1150. It is unclear whether this apparent date is merely an accidental result of what has been preserved and is known through 20th-century scholarship or whether it corresponds to some precise event or series of events.

Glass and textiles continued to be major mediums during the Seljuq period. Ceramics underwent many changes, especially in Iran, where lustre painting became widespread and where new techniques were developed for colouring pottery. Furthermore, the growth of tile decoration created a new dimension for the art of ceramics.

Inlaid metalwork became an important technique. First produced at Herāt in Iran (now in Afghanistan) in the middle of the 12th century, this type of decoration spread westward, and a series of local schools were established in various regions of the Seljuq domain. In this technique, the surfaces of utilitarian metallic objects (candlesticks, ewers, basins, kettles, and so forth) were engraved, and then silver was inlaid in the cut-out areas to make the decorative design more clearly visible.

Manuscript illustration also became an important art. Scientific books, including the medical manuals of Dioscorides and of Galen, or literary texts such as the picaresque adventures of a verbal genius known as the *Maqāmāt*, were produced with narrative illustrations throughout the text.

All of the technical novelties of the Seljuqs seem to have had one main purpose: to animate objects and books and to provide them with clearly visible and identifiable images. Even the austere art of calligraphy became occasionally animated with letters ending in human figures. The main centres for producing these arts were located in Iran and the Fertile Crescent. For reasons yet unknown, Egypt and Anatolia were far less involved. One reason may be that these two Seljuq provinces did not witness the same

Seljuq manuscript illustration.
(Left) Drawing from a manuscript of the *Maqāmāt*, 1323. In the British Museum (MS. Add 7293, f. 285v). (Right) Preparing medicine from honey, manuscript illustration from an Arabic translation of *De materia medica* of Dioscorides copied by ʿAbd Allāh ibn al-Faḍl, from Baghdad, Iraq, 13th century. In the Metropolitan Museum of Art, New York City. 12.5 cm × 17.5 cm.

By courtesy of (left) the trustees of the British Museum, and (right) the Metropolitan Museum of Art, New York, the Cora Timken Burnett collection of Persian Miniatures and other Persian art objects, bequest of Cora Timken Burnett, 1957

Urban middle-class patronage

rise of an urban middle class as did Iran, Iraq, or Syria. It would seem from a large number of art objects whose patrons are known that the main market for these works of art was the mercantile bourgeoisie of the big cities and not, as has often been believed, the princes. Seljuq decorative arts and book illustration, therefore, reflect an urban taste.

The themes and motifs used were particularly numerous. In books they tend to be illustrations of the text, even if a manuscript such as the Schefer *Maqāmāt* (1237; Bibliothèque Nationale, Paris) sought to combine a strict narrative with a fairly naturalistic panorama of contemporary life. Narrative scenes taken from books or reflecting folk stories are also common on Persian ceramics. In all mediums, however, the predominant vocabulary of images is the one provided by the older art of princes; but its meaning is no longer that of illustrating the actual life of princes but rather that of symbolizing a good and happy life. The motifs, therefore, do not have to be taken literally. Next to princely and narrative themes, there are depictions of scenes of daily life, astronomical motifs, and a myriad of topics that can be described but not understood.

While it is possible within certain limits to generalize about the subject matter of Seljuq art, regional stylistic definitions tend to be more valid. Thus the bronzes produced in northeastern Iran in the 12th century are characterized by simple decorative compositions rather than by the very elaborate ones created by the so-called school of Mosul in Iraq during the 13th century. In general, the art of metalwork exhibits a consistently growing intricacy in composition and in details to the point that individual subjects are at times lost in overlapping planes of arabesques. Ceramic pieces of Iran have usually been classified according to a more or less fictitious provenance. Kāshān ware exhibits a perfection of line in the depiction of moon-faced personages with heavily patterned clothes, while Rayy ceramic work is less sophisticated in design and execution but more vividly coloured. Sāveh and Gurgān are still other Iranian varieties of pottery. With the exception of Kāshān ware, where dynasties of ceramicists are known, all these types of Iranian pottery were contemporary with each other. In Syria, Raqqah pottery imitated Iranian ceramic wares but with a far more limited vocabulary of designs.

The Baghdad, or Arab, school

The main identifiable group of miniature painters was the so-called Baghdad school of the first half of the 13th century. The group should be called the Arab school because the subject matter and style employed could have been identified with any one of the major artistic centres of Egypt and the Fertile Crescent, and very little evidence currently exists to limit this school to one city. The miniatures painted by these artists are characterized by the colourful and often humorous way in which the urbanized Arab is depicted. The compositions, often lacking in any strong aesthetic intent, are documentary caricatures in which the artist has recorded the telling and recognizable gesture or a known and common setting or activity. In many images or compositional devices one can recognize the impact of the richer Christian Mediterranean tradition of manuscript illumination. A greater attention to aesthetic considerations is apparent in the illustrated manuscript of the Persian epic *Varqeh o-Golshāh* (Topkapı Saray Museum, Istanbul), unique in the Seljuq period.

Western Islāmic art: Moorish. The 11th to 13th centuries were not peaceful in the Maghrib. Berber dynasties overthrew each other in Morocco and the Iberian Peninsula. The Christian reconquest gradually diminished Muslim holdings in Spain and Portugal, and Tunisia was ruined during the Hilālī invasion when Bedouin tribes were sent by the Fāṭimids to prevent local independence.

Austerity of North African architecture

Two types of structures characterize the Almoravid (1056–1147) and Almohad (1130–1269) periods in Morocco and Spain. One comprises the large, severely designed Moroccan mosques such as those of Tinmel, of Ḥasan in Rabat, or of the Kutubīyah in Marrakech. They are all austere hypostyles with tall, massive, square minarets. The other distinctive type of architecture was that built for military purposes, including fortifications and, especially, massive city gates with low-slung horseshoe arches, such as the Oudaia Gate at Rabat (12th century) or the Rabat Gate at Marrakech (12th century). Palaces built in central Algeria by minor dynasties such as the Zīrids were more in the Fāṭimid tradition of Egypt than in the Almoravid and Almohad traditions of western Islām. Almost nothing is known or has been studied about North African arts other than architecture because the puritanical world of the Berber dynasties did not foster the arts of luxury.

In North Africa the artistic milieu did not change much in the 14th and 15th centuries. Hypostyle mosques such as the Great Mosque of Algiers continued to be built; *madrasahs* were constructed with more elaborate plans; the Bū ʿInānīyah *madrasah* at Fès is one of the few monumental buildings of the period. A few mausoleums were erected such as the so-called Marīnid tombs near Fès (second half of the 14th century) or the complex of Chella at Rabat (mostly 14th century). Architectural decoration in stucco or sculpted stone was usually limited to elaborate geometric patterns, epigraphic themes, and a few vegetal motifs.

A stunning exception to the austerity of North African architecture exists in Spain in the Alhambra palace com-

Almoravid and Almohad architecture in North Africa.
(Top left) The Rabat Gate, Marrakech, Mor., late 12th century, Almoravid period. (Bottom left) Interior of the Great Mosque of Algiers, Alg., 12th century, Almohad period. (Right) The square minaret of the Kutubīyah Mosque in Marrakech, Mor., Almoravid period.
(Top left) Josephine Powell, Rome, (bottom left) H. Roger-Viollet, (right) GEKS

plex at Granada. The hill site of the Alhambra had been occupied by a citadel and possibly by a palace since the 11th century, but little of these earlier constructions has remained. In the 14th century two successive princes, Yūsuf I and Muḥammad V, transformed the hill into their official residence. Outside of a number of gates built like triumphal arches and several ruined forecourts, only three parts of the palace remain intact. First there is the long Court of the Myrtles leading to the huge Hall of Ambassadors located in one of the exterior towers. This was the part of the Alhambra built by Yūsuf I. Then there is the Court of the Lions, with its celebrated lion fountain in the centre. Numerous rooms open off this court, including the elaborately decorated Hall of the Two Sisters and the Hall of the Abencerrajes. The third part, slightly earlier than the first two, is the Generalife; it is a summer residence built higher up the hill and surrounded by gardens with fountains, pavilions, and portico walks.

Importance of the Alhambra

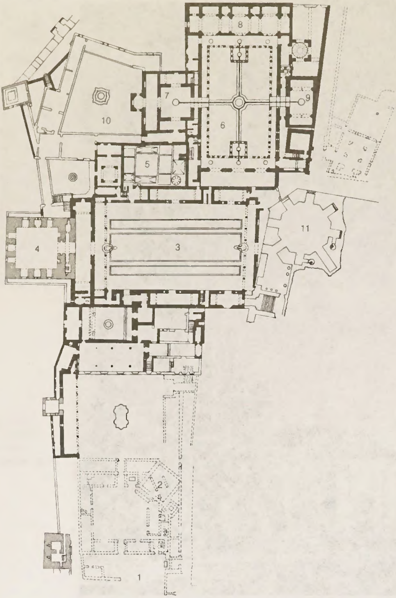
The Alhambra is especially important because it is one of the few palaces to have survived from medieval Islāmic times. It illustrates superbly a number of architectural concerns occasionally documented in literary references: the contrast between an unassuming exterior and a richly decorated interior to achieve an effect of secluded or private brilliance; the constant presence of water, either as a single, static basin or as a dynamic fountain; the inclusion of oratories and baths; the lack of an overall plan (the units are simply attached to each other).

The architectural decoration of the Alhambra was mostly of stucco. Some of it is flat, but the extraordinarily complex cupolas of *muqarnas*, such as in the Hall of the Two Sisters, appear as huge multifaceted diadems. The decoration of the Alhambra becomes a sort of paradox as well as a tour de force. Weighty, elaborately decorated ceilings, for example, are supported by frail columns or by walls pierced with many windows (light permeates almost every part of the large, domed halls). Much of the design and decoration of the Alhambra is symbolically oriented. The poems that adorn the Alhambra as calligraphic ornamentation celebrate its cupolas as domes of heaven rotating around the prince sitting under them.

Islāmic art as such ceased to be produced in Spain after 1492, when Granada, the last Moorish kingdom in Spain, fell to the Christians; but the Islāmic tradition continued in North Africa, which remained Muslim. In Morocco the so-called Sharīfian dynasties from the 16th century onward ornamentally developed the artistic forms created in the 14th century.

Most of the best known monuments of western Islāmic art are buildings, although a very original calligraphy was developed. The other arts cannot be compared in wealth and importance either with what occurred elsewhere in Islām at the same time or with earlier objects created in Spain. There are some important examples of metalwork, wood inlaid with ivory, and a lustre-glaze pottery known as Hispano-Moresque ware. The fact that the latter was made in Valencia or Málaga after the termination of Muslim rule demonstrates that Islāmic traditions in the decorative arts continued to be adhered to, if only partially. The term Mudéjar, therefore, is used to refer to all the things made in a Muslim style but under Christian rule. Numerous examples of Mudéjar art exist in ceramics and

Mudéjar and Mozarabic art



Plan of the Alhambra in Granada, Spain.
1. Alcazaba; 2. ruins of the mosque; 3. Court of the Myrtles; 4. Tower of Comares and Hall of Ambassadors; 5. baths; 6. Court of the Lions; 7. Hall of the Two Sisters; 8. Hall of the Kings; 9. Hall of the Abencerrajes; 10. Garden of Daraxa; 11. Palace of Charles V.
From G. Marcais, *L'Architecture Musulmane d'Occident*

textiles, as well as in architectural monuments such as the synagogues of Toledo and the Alcazba in Seville, where even the name of the ruling Christian prince, Don Pedro, was written in Arabic letters. The Mudéjar spirit, in fact, permeated most of Spanish architectural ornament and decorative arts for centuries, and its influence can even be found in Spanish America.

Mudéjar art must be carefully distinguished from Mozarabic art: the art of Christians under Muslim rule. Mozarabic art primarily flourished in Spain during the earlier periods of Muslim rule. Its major manifestations are architectural decorations, decorative objects, and illuminated manuscripts. Dating mostly from the 10th and 11th centuries, the celebrated illuminations for the commentary on the Revelation to John by an 8th-century Spanish abbot, Beatus of Liébana, are purely Christian subjects treated in styles possibly influenced by Muslim miniature painting or book illustration. The most celebrated example, known as the "Saint-Sever Apocalypse," is in the collection of the Bibliothèque Nationale in Paris.

Mamlūk art. The Mamlūks were originally white male slaves, chiefly Turks and Circassians from the Caucasus and Central Asia who formed the mercenary army of the various feudal states of Syria and Egypt. During the 13th century the importance of this military caste grew as the older feudal order weakened and military commanders took over power generally as nonhereditary sultans. They succeeded in arresting the Mongol onslaught in 1260 and, through a judicious but complicated system of alliance with the urban elite class, managed to maintain themselves in power in Egypt, Palestine, and Syria until 1517.

During the Mamlūk period Egypt and Syria were rich commercial emporiums. This wealth explains the quality and quantity of Mamlūk art. Most of the existing monuments in the old quarters of Cairo, Damascus, Tripoli, and Aleppo are Mamlūk; in Jerusalem almost everything visible on the Ḥaram ash-Sharīf, outside the Dome of the Rock, is Mamlūk. Museum collections of Islāmic art generally abound with Mamlūk metalwork and glass. Some of the oldest remaining carpets are Mamlūk. This creativity required, of course, more than wealth; it also required a certain will to transform wealth into art. This will was in part the desire of parvenu rulers and their cohorts to be remembered. Furthermore, architectural patronage flourished because of the institutionalization of the *waqf,* an economic system in which investments made for holy purposes were inalienable. This law allowed the wealthy to avoid confiscation of their properties at the whim of the caliph by investing their funds in religious institutions. In the Mamlūk period, therefore, there was a multiplication of *madrasah*s, *khānqāh*s, *ribāṭ*s, and *masjid*s, often with tombs of founders attached to them. The Mamlūk establishment also repaired and kept up all the institutions, religious or secular, that had been inherited by them, as can be demonstrated by the well-documented repairs carried on in Jerusalem and Damascus.

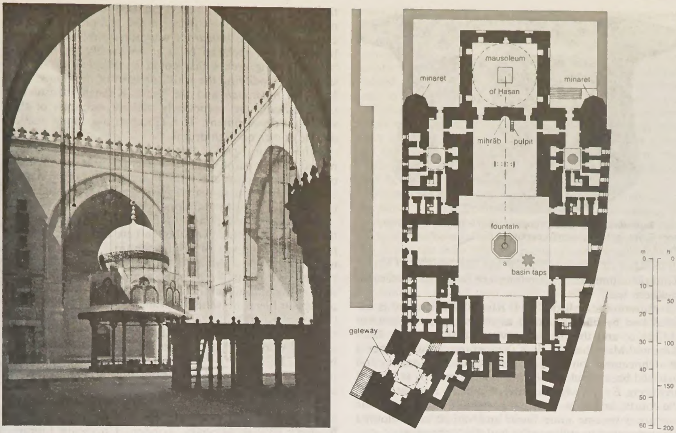
The influential role of the *waqf*

Architecture. The Mamlūks created a monumental setting for Syria and Egypt that lasted until the 20th century. It was at its most remarkable in architecture, and nearly 3,000 major monuments have been preserved or are known from texts in cities from the Euphrates to Cairo. No new architectural types came into being, although many more urban commercial buildings and private houses have been preserved than from previous centuries. The hypostyle form continued to be used for mosques and oratories, as in the Cairene mosques of Baybars I (1262–63), Nāṣir (1335), and Mu'ayyad Shaykh (1415–20). *Madrasah*s used *eyvān*s, and the justly celebrated *madrasah* of Sultan Ḥasan in Cairo (1356–62) is one of the few perfect four-*eyvān madrasah*s in the Islāmic world. Mausoleums were squares or polygons covered with domes. In other words, there were only minor modifications in the typology of architecture, and even the 15th-century buildings with interiors totally covered with ornamentation have possible prototypes in the architecture of the Seljuqs. Yet there are formal and functional features that do distinguish Mamlūk buildings. One is the tendency to build structures of different functions in a complex or cluster. Thus the Qalā'ūn mosque (1284–85) in Cairo has a mausoleum, a *madrasah,* and a hospital erected as one architectural unit. Another characteristic is the tendency of Mamlūk patrons to build their major monuments near each other. As a result, certain streets of Cairo, such as Bayn al-Qaṣrayn, became galleries of architectural masterpieces. The plans of these buildings may have had to be adapted to the exigencies of the city, but their spectacular facades and minarets competed with each other for effect. From the second half of the 14th century onward, building space for mausoleums began to be limited in Cairo, and a vast complex of commemorative monuments was created in the city's western cemetery. In Aleppo and Damascus similar phenomena can be observed.

Architectural changes in the Mamlūk period

Although Mamlūk architecture was essentially conservative in its development of building types, more originality is evident in the constructional systems used, although traditional structural features continued to be employed—*e.g.,* cupolas raised on squinches or more commonly pendentives, barrel and groin vaults, and wooden ceilings covering large areas supported by columns and piers. The main innovations are of three kinds. First, minarets became particularly elaborate and, toward the end of the period, almost absurd in their ornamentation. Facades were huge, with overwhelming portals 25 to 35 feet high.

A second characteristically Mamlūk feature was technical virtuosity in stone construction. At times this led to a superb purity of form, as in the Gate of the Cotton Merchants in Jerusalem or the complex of the Barqūq mosque in Cairo. At other times, as in the Mamlūk ar-



The madrasah of Sultan Ḥasan, Cairo, 1356–62.
(Left) Courtyard and (right) plan.
(Left) GEKS, (right) from B. Fletcher, *A History of Architecture on the Comparative Method;* Athlone Press of the University of London, London

chitecture of Baybars and Qā'it Bāy, there was an almost wild playfulness with forms. Another aspect of Mamlūk masonry was the alternation of stones of different colours to provide variations on the surfaces of buildings.

The third element of change in Mamlūk art was perhaps the most important: almost all formal artistic achievements rapidly became part of the common vocabulary of the whole culture, thus ensuring high quality of construction and decorative technique throughout the period.

With the exception of portals and *qiblah* walls, architectural decoration was usually subordinated to the architectural elements of the design. Generally the material of construction (usually stone) was carved with ornamental motifs. Stucco decoration was primarily used in early Mamlūk architecture, while coloured tile was a late decorative device that was rarely employed.

H. Roger-Viollet

Mamlūk tombs, Cairo, 14th–15th centuries.

Other arts. Like architecture, the other arts of the Mamlūk period achieved a high level of technical perfection but were often lacking in originality. The so-called "Baptistère de St. Louis" (*c.* 1310, Louvre) is the most impressive example of inlaid metalwork preserved from this period. Several Mamlūk illustrated manuscripts, such as the *Maqāmāt* (1334) in the Nationalbibliothek, Vienna, display an amazing ornamental sense in the use of colour on gold backgrounds. Mamlūk mosque lamps provide some of the finest examples of medieval glass. The wooden objects made by Mamlūk craftsmen were widely celebrated for the quality of their painted, inlaid, or carved designs. And the bold inscriptions that decorate the hundreds of remaining bronzes testify to the Mamlūks' mastery of calligraphy. None of these examples, however, exhibits much inventiveness of design.

Mongol Iran: Il-Khanid and Timurid periods. Seen from the vantage point of contemporary or later chronicles, the 13th century in Iran was a period of destructive wars and invasions. Such cities as Balkh, Nīshāpūr, or Rayy, which had been centres of Islāmic culture for nearly six centuries, were eradicated as the Mongol army swept through Iran. The turning point toward some sort of stability took place in 1295 with the accession of Maḥmūd Ghāzān to the Mongol throne. Under him and his successors (the Il-Khan dynasty), order was reestablished throughout Iran, and cities in northeastern Iran, especially Tabriz and Solṭānīyeh, became the main creative centres of the new Mongol regime. At Tabriz, for example, the Rashīdīyeh (a sort of academy of sciences and arts to which books, scholars, and ideas from all over the world were collected) was established in the early 14th century.

Existing under the Mongol rulers were a number of secondary dynasties that flourished in various provinces of Iran: the Jalāyirid dynasty, centred in Baghdad, controlled most of western Iran; the Moẓaffarid dynasty of southwestern Iran contained the cities of Isfahan, Yazd, and Shīrāz; and the Karts reigned in Khorāsān. Until the last decade of the 14th century, however, all the major cultural centres were in western Iran. Under Timur (1336–1405; the Timurid dynasty) and his successors, however, northeastern Iran, especially the cities of Samarkand and Herāt, became focal points of artistic and intellectual activity. But Timurid culture affected the whole of Iran either directly or through minor local dynasties. Many

The "Baptistère de St. Louis," copper inlaid with gold and silver, from Egypt, c. 1310, Mamlūk period. In the Louvre, Paris. Ht. 22.8 cm.

By courtesy of the Musee du Louvre, Paris; photograph, Cliche Musees Nationaux, Paris

Timurid monuments, therefore, are found in western or southern Iran.

Architecture. Stylistically, Il-Khanid architecture is defined best by buildings such as the mosque of Varāmīn (1322–26) and the mausoleums at Sarakhs, Merv, Rād-Kān, and Marāgheh. In all of these examples, the elements of architectural composition, decoration, and construction that had been developed earlier were refined by Il-Khanid architects. *Eyvān*s were shallower but better integrated with the courts; facades were more thoughtfully composed; the *muqarnas* became more linear and varied; and coloured tiles were used to enhance the building's character.

Il-Khanid refinement and taste for the grandiose

The architectural masterpiece of the Il-Khanid period is the mausoleum of Öljeitü at Solṭānīyeh. With its double system of galleries, eight minarets, large blue-tiled dome, and an interior measuring 80 feet (25 metres), it is clear that the building was intended to be imposing. Il-Khanid attention to impressiveness of scale also accounted for the ʿAlī Shāh mosque in Tabriz, whose *eyvān* measuring 150 by 80 by 100 feet (45 by 25 by 30 metres) was meant to be the largest ever built. The *eyvān* vault collapsed almost immediately after it had been constructed, but its walls, 35 feet (10 metres) thick, remain as a symbol of the grandiose taste of the Il-Khanids. In the regions of Isfahan and Yazd numerous smaller mosques (often with unusual plans) and less pretentious mausoleums, as well as palaces with elaborate gardens, were built in the 14th century. These buildings were constructed to provide a monumental setting for the Islāmic faith and for the authority of the state. The study of these buildings began only in the mid-20th century, and therefore no definitive conclusions have been reached as to whether regional or pan-Iranian stylistic and formal features predominated.

The Timurid period began architecturally in 1390 with the sanctuary of Aḥmad Yasavī in Turkistan. Between 1390 and the last works of Sultan Ḥusayn Bayqara almost a century later, hundreds of buildings were constructed at Herāt, many of which have been preserved, although few have been studied except by Central Asian scholars. The most spectacular examples of Timurid architecture are found in Samarkand, Herāt, Meshed, Khargird, Tayābād, Baku, and Tabriz, although important Timurid structures were also erected in southern Iran.

Architectural projects were well patronized by the Timurids as a means to commemorate their respective reigns. Every ruler or local governor constructed his own sanctuaries, mosques, and, especially, memorial buildings dedicated to holy men of the past. While the Shāh-e Zendah in Samarkand—a long street of mausoleums comparable to the Mamlūk cemetery of Cairo—is perhaps the most accessible of the sites of Timurid commemorative architecture, more spectacular ones are to be seen at Meshed, Torbat-e Sheykh Jām, and Mazār-e Sharīf. The Timurid princes also erected mausoleums for themselves, such as the Gūr-e Amīr and the ʿIshrat-Khāneh in Samarkand.

Timurid concern with a commemorative architecture

Major Timurid buildings, such as the so-called mosque of Bībī Khānom, the Gūr-e Amīr mausoleum, the mosque of Gowhar Shād in Meshed, or the *madrasah*s at Khargird and Herāt, are all characterized by strong axial symmetry. Often the facade on the inner court repeats the design of the outer facade, and minarets are used to frame the composition. Changes took place in the technique of dome construction. The *muqarnas* was not entirely abandoned but was often replaced by a geometrically rigorous net of intersecting arches that could be adapted to various shapes by modifying the width or span of the dome. The Khargird *madrasah* and the ʿIshrat-Khāneh mausoleum in Samarkand are particularly striking examples of this structural development. The Timurids also made use of double domes on high drums.

In the Timurid period the use of colour in architecture reached a high point. Every architectural unit was divided, on both the exterior and interior, into panels of brilliantly coloured tiles that sometimes were mixed with stucco or terra-cotta architectural decorations.

Painting. A new period of Persian painting began in the Mongol era, and, even though here and there one can recognize the impact of Seljuq painting, on the whole it is a limited one. Although the new style was primarily expressed in miniature painting, it is known from literary sources that mural painting flourished as well. Masterpieces of Persian literature were illustrated: first the *Shāh-nāmeh* ("Book of Kings") by the 11th-century poet Ferdowsī and then, from the second half of the 14th century, lyrical and mystical works, primarily those by the 12th-century poet Neẓāmī. Historical texts or chronicles such as the *Jāmiʿ at-tawārīkh* ("Universal History of Rashīd ad-Dīn") were also illustrated, especially in the early Mongol period.

The first major monument of Persian painting in the Mongol period is a group of manuscripts of the *Jāmiʿ at-tawārīkh* (British Museum, London; University Library, Edinburgh; and Topkapı Saray Museum, Istanbul). The miniatures are historical narrative scenes. Stylistically they are related to Chinese painting—an influence introduced by the Mongols during the Il-Khanid period.

Chinese influence can still be discovered in the masterpiece of 14th-century Persian painting, the so-called Demotte *Shāh-nāmeh.* Illustrated between 1320 and 1360, its 56 preserved miniatures have been dispersed all over the world. The compositional complexity of these paintings can be attributed to the fact that several painters probably were involved in the illustration of this manuscript and that these artists drew from a wide variety of different stylistic sources (*e.g.,* Chinese, European, local Iranian traditions). Its main importance lies in the fact that it is the earliest known illustrative work that sought to depict in a strikingly dramatic fashion the meaning of the Iranian epic. Its battle scenes, its descriptions of fights with monsters, its enthronement scenes are all powerful representations of the colourful and often cruel legend of Iranian kingship. The artists also tried to express the powerlessness of man confronted by fate in a series of mourning and death scenes.

The expressive importance of the *Shāh-nāmeh* miniatures

The Demotte *Shāh-nāmeh* is but the most remarkable of a whole series of 14th-century manuscripts, all of which suggest an art of painting in search of a coherent style. At the very end of the period a manuscript such as that of the poems of Sultan Aḥmad (Freer Gallery of Art, Washington, D.C.) still exhibits an effective variety of established themes, while some of the miniatures in the Deutsche Staatsbibliothek, Berlin, and in the Topkapı Saray, Istanbul, illustrate the astounding variety of styles studied or copied by Persian masters.

A more organized and stylistically coherent period in Persian painting began around 1396 with the Khwāju Kermānī manuscript (British Museum) and culminated between 1420 and 1440 in the paintings produced by the Herāt school, where the emperor Baysunqur created an academy in which classical Iranian literature was codified, copied, and illustrated. Although several *Shāh-nāmeh*s are known from this time, the mood of these manuscripts is no longer epic but lyrical. Puppet-like figures almost unemotionally engage in a variety of activities always set in an idealized garden or palace depicted against a rich gold background. It is a world of sensuous pleasure that also embodies the themes of a mystically interpreted lyrical poetry, for what is represented is not the real world but a divine paradise in the guise of a royal palace or garden. These miniatures easily became clichés, for later artists

Mausoleum of Öljeitü at Solṭānīyeh, 1305–13, Il-Khanid period.
Josephine Powell, Rome

endlessly repeated stereotyped formulas. But at its best, as in the Metropolitan Museum Neẓāmī, this style of Persian painting succeeds in defining something more than mere ornamental colourfulness. It expresses in its controlled lyricism a fascinating search for the divine, similar to the search of such epic characters as Neẓāmī, Rūmī, or Ḥāfeẓ—at times earthly and vulgar, at other times quite ambiguous and hermetic, but often providing a language for the ways in which human beings can talk about God.

Behzād and his school

Another major change in Persian painting occurred during the second half of the 15th century at Herāt under Ḥusayn Bayqara. This change is associated with the first major painter of Islāmic art, Behzād. Many problems of attribution are still posed about Behzād's art, and, in the examples that follow, works by his school, as well as images by the master's own hand, are included. In the Garrett *Ẓafar-nāmeh* (*c.* 1490), the Egyptian Cairo National Library's Būstān (1488), or the British Museum's Neẓāmī (1493–94), the stereotyped formulas of the earlier lyric style were endowed with new vitality. Behzād's interest in observing his environment resulted in the introduction of more realistic poses and the introduction of numerous details of daily life or genre elements. His works also reflect a concern for a psychological interpretation of the scenes and events depicted. It is thus not by chance that portraits have been attributed to Behzād.

Persian art of the Mongol period differs in a very important way from any of the other traditions of the middle period of Islāmic art. Even though Iran, like all other areas at that time, was not ethnically homogeneous, its art tended to be uniquely "national." In architecture nationalism was mostly a matter of function, for during this period the Shīʿites grew in importance, and new monumental settings were required for their holy places. Iranian individualism is especially apparent in painting, in which Chinese and other foreign styles were consistently adapted to express intensely Iranian subjects, thereby creating a uniquely Persian style.

LATE PERIOD

The last period of an Islāmic artistic expression created within a context of political and intellectual independence was centred in the Ottoman, Ṣafavid, and Mughal empires. Although culturally very different from each other, these three imperial states shared a common past, a common consciousness of the nature of their ancestry and of the artistic forms associated with it. Painters and architects moved from one empire to the other, especially from Iran to India; Ottoman princes wrote Persian poetry, and Ṣafavid rulers spoke Turkish. But most of all, they were aware of the fact that they were much closer to each other than to any non-Islāmic cultural entity. However different their individual artistic forms may have been, they collected each other's works, exchanged gifts, and felt that they belonged to the same world.

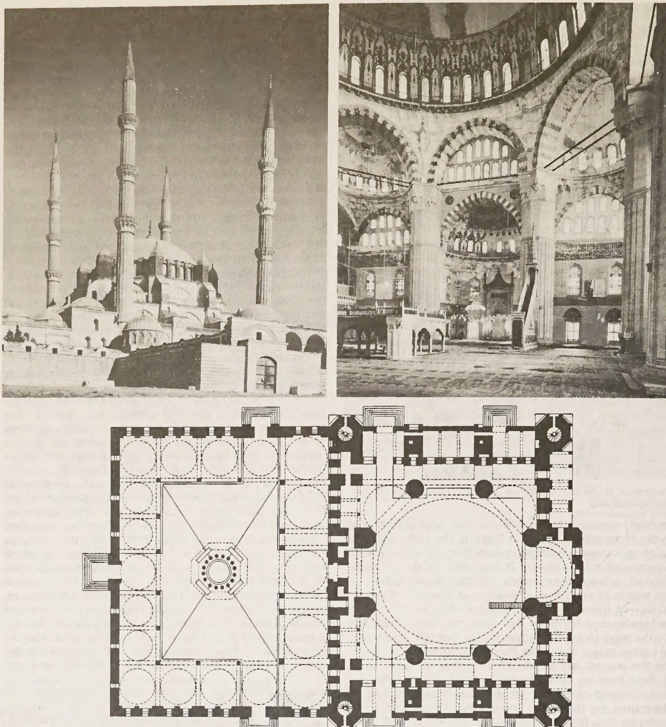
Ottoman art. The Ottomans were originally only one of the small Turkmen principalities (*beylik*s) that sprang up in Anatolia around 1300 after the collapse of Seljuq rule. In many ways, all the *beylik*s shared the same culture, but it was the extraordinary political and social attributes of the Ottomans that led them eventually to swallow up the other kingdoms, to conquer the Balkans, to take Constantinople in 1453, and to control almost the whole of the Arab world by 1520. Only in the 19th century did this complex empire begin to crumble. Thus, while Ottoman art, especially architecture, is best known through the monuments in Turkey, there is, in fact, evidence of Ottoman art extending from Algiers to Cairo in North Africa, to Damascus in the Levant, and in the Balkans from Sarajevo, Yugos., to Sofia, Bulg.

Architecture. The grand tradition of Ottoman architecture, established in the 16th century, was derived from two main sources. One was the rather complex development of new architectural forms that occurred all over Anatolia, especially at Manisa, İznik, Bursa, and Selçuk in the 14th and early 15th centuries. In addition to the usual mosques, mausoleums, and *madrasah*s, a number of buildings called *tekke*s were constructed to house dervishes (members of mystical fraternities) and other holy men who lived communally. The *tekke* (or *zeviye*) was often joined to a mosque or mausoleum. The entire complex was then called a *külliye.* All these buildings continued to develop the domed, central-plan structure, constructed by the Seljuqs in Anatolia. The other source of Ottoman architecture is Christian art. The Byzantine tradition, especially as embodied in Hagia Sophia, became a major source of inspiration. Byzantine influence appears in such features as stone and brick used together or in the use of pendentive dome construction. Also artistically influential were the contacts that the early Ottomans had with

Origin of the *tekke* and *külliye*

By courtesy of the Smithsonian Institution, Freer Gallery of Art, Washington, D.C.

Mourning scene at the bier of Alexander the Great, miniature from the Demotte *Shāh-nāmeh* ("Book of Kings") of Ferdowsī, colour and gold on paper, Tabriz school, 14th century. In the Freer Gallery of Art, Washington, D.C. 25 cm × 28 cm.



Selim Mosque at Edirne, Tur., designed by Sinan, 1569–75. (Top left) Exterior; (top right) interior; (bottom) plan.

(Top) K. Scholz—Shostal/EB Inc., (bottom) from G. Goodwin, *A History of Ottoman Architecture;* Johns Hopkins University Press

Italy. Thus, in several mosques at Bursa, Tur., there are stylistic parallels in the designs of the exterior facade and of windows, gates, and roofs to features found in Italian architecture. A distinctive feature of Ottoman architecture is that it drew from both Islāmic and European artistic traditions and was, therefore, a part of both.

The apogee of Ottoman architecture was achieved in the great series of *külliye*s and mosques that still dominate the Istanbul skyline: the Fatih *külliye* (1463–70), the Bayezid Mosque (after 1491), the Selim Mosque (1522), the Şehzade *külliye* (1548), and the Süleyman *külliye* (after 1550). The Şehzade and Süleyman *külliye*s were built by Sinan, the greatest Ottoman architect, whose masterpiece is the Selim Mosque at Edirne, Tur. (1569–75). All of these buildings exhibit total clarity and logic in both plan and elevation; every part has been considered in relation to the whole, and each architectural element has acquired a hierarchic function in the total composition. Whatever is unnecessary has been eliminated. This simplicity of design in the late 15th and 16th centuries has often been attributed to the fact that Sinan and many Ottoman architects were first trained as military engineers. Everything in these buildings was subordinated to an imposing central dome. A sort of cascade of descending half domes, vaults, and ascending buttresses leads the eye up and down the building's exterior. Minarets, slender and numerous, frame the exterior composition, while the open space of the surrounding courts prevents the building from being swallowed by the surrounding city. These masterpieces of Ottoman architecture seem to be the final perfection of two great traditions: a stylistic and aesthetic tradition that had been indigenous to Istanbul since the construction of the Byzantine church of Hagia Sophia in the 6th century and the other Islāmic tradition of domical construction dating to the 10th century.

The tragedy of Ottoman architecture is that it never managed to renew its 16th-century brilliance. Later buildings, such as the impressive Sultan Ahmed mosque in Istanbul, were mostly variations on Sinan's architecture, and sometimes there were revivals of older building types, especially in the provinces. Occasionally, as in the early 18th-century Nûruosman mosque in Istanbul, interesting new variants appear illustrating the little-known Turkish Baroque style. The latter, however, is more visible in ornamental details or in smaller buildings, especially the numerous fountains built in Istanbul in the 18th century.

The sources of the Turkish Baroque are probably to be sought in the Baroque architecture of Vienna and the bordering Austro-Hungarian states. Throughout the 18th and 19th centuries, a consistent Europeanization of a local tradition occurs in the Ottoman empire.

While mosques and *külliyes* are the most characteristic monuments of Ottoman architecture, important secular buildings were also built: baths, caravansaries, and especially the huge palace complex of Topkapı Saray at Istanbul, in which 300 years of royal architecture are preserved in its elaborate pavilions, halls, and fountains.

Other arts. Architectural decoration was generally subordinated to the structural forms or architectonic features of the building. A wide variety of themes and techniques originating from many different sources were used. One decorative device, the Ottoman version of colour-tile decoration, deserves particular mention, for it succeeds in transforming smaller buildings such as the mosque of Rüstem Paşa in Istanbul into a visual spectacle of brilliant colours. The history and development of this type of ceramic decoration is intimately tied to the complex and much controverted problem of the growth of several distinctive Ottoman schools of pottery: İznik, Rhodian, and Damascus ware. Both in technique and in design, Ottoman ceramics are the only major examples of pottery produced in the late Islāmic period.

Widespread use of tile decoration

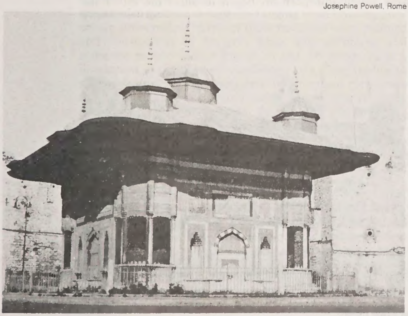
Ottoman miniature painting does not compare in quality with Persian painting, which originally influenced the Turkish school. Yet Ottoman miniatures do have a character of their own, either in the almost folk art effect of religious images or in the precise depictions of such daily events as military expeditions or great festivals. Among the finest examples of the latter is the manuscript *Surname-i Vehbi* (Topkapı Saray Museum, Istanbul) painted by Levnî in the early 18th century.

The production of metalwork, wood inlaid with ivory, Usak carpets, and textiles flourished under the Ottomans, both in Istanbul workshops sponsored by the sultan and in numerous provincial centres. The influence of these ornamental objects on European decorative arts from the 16th through the 19th century was considerable.

Ṣafavid art. The Ṣafavid dynasty was founded by Esmā'īl I (1501–24). The art of this dynasty reached its zenith during the reigns of Ṭahmāsp (1524–76) and of 'Abbās I (1588–1629). This phase of the Ṣafavid period also marked the last significant development of Islāmic art in Iran, for after the middle of the 17th century original creativity disappeared in all mediums. Rugs and objects in silver, gold, and enamel continued to be made and exhibited a considerable technical virtuosity, even when they were lacking in inventiveness.

The Ṣafavids abandoned Central Asia and northeastern Iran to a new Uzbek dynasty that maintained the Timurid style in many buildings (especially at Bukhara) and briefly sponsored a minor and derivative school of painting. Only the great sanctuary of Meshed was being kept up and built-up, but, like many of the other religious sanctuaries of the time—Qom, an-Najaf, Karbalā', it is still far too little known to lend itself to coherent analysis. For this is the time when Shī'ism became a state religion and for the first time in Islām there appeared an organized ecclesiastical system rather than the more or less loose spiritual and practical leadership of old. The main centres of the Ṣafavid empire were Tabriz and Ardabīl in the northwest, with Kazvin in the central region, and, especially, Isfahan in the west. The Ṣafavid period, like the Ottoman era, was an imperial age, and therefore there is hardly a part of Iran where either Ṣafavid buildings or major Ṣafavid restorations cannot be found. The dynasty spent much money and effort on the building of bridges, roads, and caravansaries to encourage trade.

Josephine Powell, Rome



Turkish Baroque style exemplified by the Fountain of Ahmed III, Istanbul, 1728.

Architecture. The best known Ṣafavid monuments are located at Isfahan, where 'Abbās I built a whole new city. According to one description, it contained 162 mosques, 48 *madrasahs*, 1,802 commercial buildings, and 283 baths. Most of these buildings no longer survive, but what has remained constitutes some of the finest monuments of Islāmic architecture.

At the centre of Isfahan is the Meydān-e Shāh, a large open space, about 1,670 by 520 feet (510 by 158 metres), originally surrounded by trees. Used for polo games and parades, it could be illuminated with 50,000 lamps. Each side of the *meydān* was provided with the monumental facade of a building. On one of the smaller sides was the entrance to a large mosque, the celebrated Masjed-e Shāh. On the other side was the entrance into the bazaar or marketplace. On the longer sides were the small funerary mosque of Sheykh Loṭfollāh and, facing it, the 'Alī Qāpū, the "high gate," the first unit of a succession of palaces and gardens that extended beyond the *meydān*, most of which have now disappeared except for the Chehel Sotūn, the palace of the "Forty Columns." The 'Alī Qāpū was, in its lower floors, a semipublic place to which petitions could be brought, while its upper floors are a world of pure fantasy—a succession of rooms, halls, and balconies overlooking the city, which were purely for the prince's pleasure.

The Meydān-e Shāh

The Meydān-e Shāh unites in a single composition all the concerns of medieval Islāmic architecture: prayer, commemoration, princely pleasure, trade, and spatial effect. None of the hundreds of other remaining Ṣafavid monuments can match its historical importance, and in it also are found the major traits of Ṣafavid construction and decoration. The forms are traditional, for the most part, and even in vaulting techniques and the use of coloured tiles it is to Timurid art that the Ṣafavids looked for their models. The Persian architects of the early 17th century sought to achieve a monumentality in exterior spatial composition (an interesting parallel to the interior spaciousness created at the same time by the Ottomans); a logical precision in vaulting that was successful in the Masjed-e Shāh but rapidly led to cheap effects or to stucco imitations; and a coloristic brilliance that has made the domes and portals of Isfahan justly famous.

Painting. In the 16th and 17th centuries, possibly for the first time in Islāmic art, painters were conscious of historical styles—even self-conscious. Miniatures from the past were collected, copied, and imitated. Patronage, however, was fickle. A royal whim would gather painters together or exile them. Many names of painters have been preserved, and there is little doubt that the whim of patrons was being countered by the artists' will to be socially and economically independent as well as individually recognized for their artistic talents. Too many different impulses, therefore, existed in Ṣafavid Iran for painting to follow any clear line of development.

Importance of historical styles of painting

Three major painting styles, or schools (excluding a number of interesting provincial schools), existed in the Ṣafavid period. One school of miniature painting is exemplified by such masterpieces as the Houghton *Shāhnāmeh* (completed in 1537), the Jāmī *Haft owrang* (1556–1665; Freer Gallery of Art, Washington, D.C.), or the illustrations to stories from Ḥāfeẓ which have not been identified in detail (Fogg Art Museum, Cambridge, Mass., and in a private collection). However different they are from each other, these large, colourful miniatures all were executed in a grand manner. Their compositions are complex, individual faces appear in crowded masses, there is much diversification in landscape, and, despite a few ferocious details of monsters or of strongly caricaturized poses and expressions, these book illustrations are concerned with an idealized vision of life. The sources of this school lie with the Timurid academy. Behzād, Sulṭān Muḥammad, Sheykhzādeh, Mīr Sayyid ʿAlī, Āqā Mīrak, and Maḥmūd Muṣavvīr continued and modified, each in his own way, the ideal of a balance between an overall composition and precise rendering of details.

The miniatures of the second tradition of Ṣafavid painting seem at first to be like a detail out of the work of the previously discussed school. The same purity of colour, elegance of poses, interest in details, and assertion of the individual figure is found. Āqā Reẓā and Reẓā ʿAbbāsī (both active around 1600) excelled in these extraordinary portrayals of poets, musicians, courtiers, and aristocratic life in general.

Portraiture and genre painting

In both traditions of painting, the beautiful personages depicted frequently are satirized; this note of satirical criticism is even more pronounced in portraiture of the time. But it is in pen or brush drawings, mostly dating from the 17th century, that the third aspect of Ṣafavid painting appeared: an interest in genre, or the depiction of minor events of daily life (*e.g.*, a washerwoman at work, a tailor sewing, an animal). With stunning precision Ṣafavid artists showed a whole society falling apart with a cruel sympathy totally absent from the literary documents of the time.

While architecture and painting were the main artistic vehicles of the Ṣafavids, the making of textiles and carpets was also of great importance. It is in the 16th century that a hitherto primarily nomadic and folk medium of the decorative arts was transformed into an expression of royal and urban tasks by the creation of court workshops. The predominantly geometric themes of earlier Iranian carpets were not abandoned entirely but tended to be replaced by vegetal, animal, and even occasional human motifs. Great schools of carpetmaking developed particularly at Tabriz, Kāshān, and Kermān.

Mughal art. Since the culture of the Mughals was intimately connected to the indigenous Hindu traditions of the Indian subcontinent, their art will be treated only synoptically in this article. (For a more detailed account, the reader should see the section on Mughal art in the visual arts portion of the article SOUTH ASIAN ARTS).

The art of the Mughals was similar to that of the Ottomans in that it was a late imperial art of Muslim princes. Both styles were rooted in several centuries (at least from the 13th century onward) of adaptation of Islāmic functions to indigenous forms. It was in the 14th-century architecture of South Asian sites such as Tughluqābād, Gaur, and Ahmadābād that a uniquely Indian type of Islāmic hypostyle mosque was created, with a triple axial nave, corner towers, axial minarets, and cupolas. It was also during these centuries that the first mausoleums set in scenically spectacular locations were built. By then the conquering Muslims had fully learned how to utilize local methods of construction, and they adapted South Asian decorative techniques and motifs.

Mughal art was in continuous contact with Iran or, rather, with the Timurid world of the second half of the 15th century. The models and the memories were in Herāt or Samarkand, but the artists were raided from Ṣafavid Iran, and the continuous flow of painters from Iran to the Mughal empire is a key factor in understanding Mughal painting.

The mausoleum of Humāyūn in Delhi (1565–69), the city of Fateḥpūr Sīkri (from 1569 onward), and the Tāj Mahal at Āgra (1631–53) summarize the development of Mughal architecture. In all three examples it can be seen that what Mughal architecture brought to the Islāmic tradition (other than traditional Indian themes, especially in decoration) was technical perfection in the use of red sandstone or marble as building and decorative materials.

Mughal painting

In Mughal painting the kind of subject that tended to be illustrated was remarkably close to those used in Ṣafavid history books—legendary stories, local events, portraits, genre scenes. What evolved quickly was a new manner of execution, and this style can be seen as early as about 1567, when the celebrated manuscript *Dāstān-e Amīr Ḥamzeh* ("Stories of Amīr Ḥamzeh") was painted (some 200 miniatures remain and are found in most major collections of Indian miniatures, especially at the Freer Gallery of Art, Washington, D.C.). Traditional Iranian themes—battles, receptions, feasts—acquired monumentality, not only because of the inordinate size of the images but also because almost all of the objects and figures depicted were seen in terms of mass rather than line. Something of the colourfulness of Iranian painting was lost, but instead images acquired a greater expressive power. Mughal portraiture gave more of a sense of the individual than did the portraits of the Ṣafavids. As in a celebrated representation of a dying courtier in the Boston Museum of Fine Arts, Mughal drawings could be poignantly naturalistic. Mood was important to the Mughal artist—in many paintings of animals there is a playful mood; a sensuous mood is evident in the first Muslim images to glorify the female body and the erotic.

In summary it can be said that the Mughals produced an art of extraordinary stylistic contrasts that reflected the complexities of its origins and of its aristocratic patronage.

ISLĀMIC ART UNDER EUROPEAN INFLUENCE AND CONTEMPORARY TRENDS

It is extremely difficult to decide when, how, and to what extent European art began to affect the art of the

Inge Morath—Magnum

The Meydān-e Shāh, originally built as a polo ground by Shāh ʿAbbās I the Great (reign 1588–1629), at Isfahan, Iran. Facing the square on the left is the mosque of Sheykh Loṭfollāh, in the centre the Masjed-e Shāh, and at the right the palace of ʿAlī Qāpū.

traditional Muslim world. Ottoman architecture was from the beginning affected by Western influences. In Mughal India, European landscapes and Western spatial concerns influenced painting in the 18th century; and Persian painting has exhibited constant Western influence since the 17th century. Thus, Islāmic art began to be affected by European traditions before Europe began (in the 18th and 19th centuries) its conquests of most of the Muslim world. Since the Ottomans ruled North Africa (except Morocco), Egypt, Syria, Palestine, as well as the Balkans, much of the Muslim world was first introduced to "modern" European art through its adaptation in Istanbul or in other major Ottoman cities like Smyrna or Alexandria.

European influence tended to have been mostly limited to architecture. Nineteenth-century European engineers and architects, for example, adapted modern structural technology and decorative styles to local Islāmic needs or idioms: the Sūq al-Ḥamīdīyah bazaar in Damascus was built with steel roofing; the Hejaz railway station at Damascus was decorated in a sort of Oriental Art Nouveau style.

Revival of the decorative arts

During actual European occupation of Muslim territory, there was a conscious revival of traditional decorative arts, but new techniques were often employed. This especially occurred in India and Morocco, where the retail success of an art object depended less on the local tradition than on the taste of the Europeans. What was romantic to a European, therefore, was no longer part of the world of the newly enriched and Europeanized Muslim. Much of the Europeanized architecture was drab and pretentious. The only real artistic accomplishment of this period was in the preservation and encouragement of the traditional techniques and designs of the decorative arts. The latter often had to be maintained artificially through government subsidies, for the local market, except in Morocco or India, was more easily seduced by second-rate European objects.

During the period of occupation it was questioned whether alien techniques necessarily brought with them new forms. This mood was clearly expressed in literature but less so in the visual arts, since the quality of Muslim art had deteriorated so much in the decades preceding European arrival that there was no longer a lively creative force to maintain. As various schools based on the École des Beaux-Arts in Paris were formed, however, the faculties and the students suffered from constant uncertainty as to whether they should preserve an art that was mostly artisanal or revolutionize it altogether.

20th-century currents

It is much more difficult to define in broad terms the characteristics of art in Muslim countries after the formation of independent countries in the 1940s and '50s. Extensive planning programs and building projects have been undertaken in even the poorest countries; and the wealthy Arab states, as well as pre-revolutionary Iran, transformed their traditional cities and countryside with spectacular modern complexes ranging from housing projects to universities. Many of these buildings were planned and constructed by Western firms and architects, and some are mere copies of European and American models, ill-adapted to the physical conditions and visual traditions of the Muslim world. Others are interesting and even sensitive projects: spectacular and technically innovative, such as the Intercontinental Hotel in Mecca (Frei and Otto) and the Haj Terminal of the King Abdul Aziz International Airport at Jidda, Saudi Arabia (the U.S. firm of Skidmore, Owings & Merrill); or intelligent and imaginative, such as the government buildings of Dhākā, Bangladesh (designed by the late Louis Kahn of the United States), or in the numerous buildings designed by the Frenchman André Ravereau in Mali or Algeria. Furthermore, within the Muslim world emerged several schools of architects that adopted modes of an international language to suit local conditions. The oldest of these schools are in Turkey, where architects such as Eldhem and Cansever, among many others, built highly successful works of art. Other major Muslim contributors to a contemporary Islāmic architecture are the Iranians Nader Ardalan and Kemzan Diba, the Iraqis Rifat Chaderji and Muhammad Makkiya, the Jordanian Rassem Badran, or the Bangladeshi Mazhar ul-Islam. Finally, a unique message was being transmitted by the visionary Egyptian architect Hassan Fathy, who, in eloquent and prophetic terms, urged that the traditional forms and techniques of vernacular architecture be studied and adapted to contemporary needs. Directly or indirectly, his work has inspired many young architects in the Muslim world and has led to a host of fascinating private houses, mosques, and educational facilities. The Aga Khan Award for Architecture was instituted to encourage genuine and contemporary architectural innovation in Muslim lands.

The results of dozens of new art schools and of a more enlightened patronage than during the 19th century are perhaps less spectacular in the other arts, and especially in painting. In spite of several interesting attempts to deal with calligraphy, with geometric designs, or with local folk arts, successes so far have not been clearly identified. But Turkey, Jordan, Egypt, Morocco, Iraq, Pakistan, and Indonesia all have produced talented artists.

EVALUATION

In order to evaluate and to understand a millenary artistic tradition spread over an area extending from Spain to India, the emphasis of this article has had to be on those features that relate the monuments to each other rather than on the myriad of characteristics that differentiate them. A few words about the latter are essential, however, for very soon after the formation of Islāmic culture (certainly by 1000), it seems clear that the nature of aesthetic impulses and of visual expectations began to vary. The question is one of determining what may be called the break-off points: the areas, moments, or forces that led to differentiations. One such point is the early 14th century, for almost everywhere in Islām artistic functions, forms, and techniques were renewed. And it is quite easy to separate the arts that followed the turn of the century from those that preceded it.

Importance of Iranian artistic traditions

Next to this chronological break-off, there are cultural ones, one might almost say ethnic ones, even though their ethnic association is often debatable. The clearest instance is that of Iran, whose artists and craftsmen, almost from the time of the first groups of Nīshāpūr ceramics, used distinctive techniques, styles, and especially subjects, many of which can be traced to pre-Islāmic times. The existence of a forceful Iranian personality in Islāmic art is self-evident, and its impact is found in almost all other subdivisions of the culture. Although it was not a single or even (until the 16th century) a politically or socially unified personality, it found uniqueness, possibly because it soon became (as early as in the 9th century) strongly conscious of its ancient past. The fact of that consciousness seems more important than the individual and on the whole scarce motifs it picked up from the past. A more curious example is that of the Ottomans and of the Arabs. For their ethnic past, in Central Asia and Arabia, respectively, played only a minor part in the formation of their art and was often intellectually rejected. At the same time and with notable exceptions, neither entity consistently sought models and ideas in the pre-Islāmic art of the region they had occupied. If they succeeded in creating an original artistic expression, it is in large part because of their success in creating a viable social order: the Ottoman imperial system of the 15th century, the urban order supported by military feudalism of Egypt, Syria, and North Africa. In these areas it is less a land than a society that provided the visual arts with their own distinctiveness, and it is only in recent years that Ottoman art began to be seen as Turkish and Mamlūk art as Arab. The case of India lies somewhere between the Iranian and Ottoman instances. Created by an imperial overlay on a powerful alien culture, it never entirely escaped the forms of the latter.

Thus, one can distinguish the following large cultural entities within Islāmic art: Ottoman, western Islāmic, Egypt and Fertile Crescent, Iran, India. They were all distinctive by the early 14th century. Detailed studies, of course, manage to find many additional subdivisions in time and space, and much mid-20th-century scholarship tended to work in those directions.

The unity of functions

Among the features that appear to unite these various traditions and especially to separate them collectively from other large artistic and cultural units is the unity of functions. There was created, in other words, an Islāmic religious and social function that is unique to Muslim lands. It was a diversified function, and its monuments are not alike in their forms. But they are alike in the human activity for which they were built. Limited in symbolic forms (*miḥrāb*, minaret, calligraphy as decoration), the Muslim function could be adapted to any architectural or ornamental tradition; and it was, not only in the cultures examined above, but in China, Indonesia, Africa—wherever Islām spread. The key concept here is that of a community of attitudes and of the uses of forms rather than of the making of forms.

There is a corollary to this conclusion that leads to the second level of an attempt to identify Islāmic visual arts as a whole; namely, that, as Islām limited its system of religious visual symbols, it developed a set of secular values. From the very beginning there occurred a major art of trade and of the city, as well as an art of the palace. More than any other culture and certainly earlier than any other, the Muslim world created a number of secular tastes and sponsored techniques of secular beautification. The result lies, on the one hand, in a striking succession of palaces from Khirbat al-Mafjar to the Alhambra or to Fatehpur Sīkri. It lies also in the impetus given to techniques of ceramics, textiles, and metalwork. These all tended to be the techniques of the artisan, and their importance lies not so much in the manufacture of an occasional object of art as in the raising of the level of quality of all industrial or decorative arts. This particular feature of the Islāmic tradition survived all political misfortunes. Remarkably beautiful objects were made as late as the early 19th century, and the techniques and traditions have often been revived in the 20th century with considerable success. Historically, Islāmic art became a sort of secular consciousness of artistic traditions elsewhere. Renaissance madonnas, for instance, were provided sometimes with halos containing Arabic inscriptions; bodies of saints were buried in Muslim cloth; Christian princes collected objects of Islāmic art; and *turquerie,* or Turkish themes, lay behind one of the styles of European decorative arts in the Baroque period of the 17th and 18th centuries. All this was possible also because the themes of Islāmic art almost never possessed the specificity of meaning that would make them unsuitable for use by others. Ambiguous in their abstraction of subjects and of styles, works of Islāmic art tended at times to the facile multiplication of known formulas. Yet again at this level, it was the user who determined the value of the form used.

The Islāmic visual vocabulary

All this is not to say that Islāmic art did not develop an internal visual vocabulary with a depth of its own. From the mosaics of the Umayyad Mosque of Damascus to the Alhambra or to certain Persian ceramics, one can determine the existence of concrete symbolic systems, royal and religious. It is even possible to see in the abstract arabesque or in certain uses of calligraphy attempts to express an early Muslim vision of the divine, while the glorious colour of Iranian mosques may reflect the more complex mystical thought of Shīʿism. There is no doubt that further research will provide many more examples of a meaningful visual symbolic system in the Muslim world. But in most instances that have already been studied—in particular Umayyad and Seljuq art—the remarkable point has been that such symbols did not last and that they were soon misunderstood or ignored. This refusal to be committed to visual symbols is reflected in the little that is known about Islāmic writing on art. It is only very incidentally that references are made to the value or meaning of visual expression; there are no theories on art, and even the religious injunctions against representations are a minute and almost incidental aspect of religious literature. Much more is known about individuals—ceramicists and metalworkers in early times, painters and architects in later times. The emphasis has always been on their technical skill, on their ability to do visual tricks, or on the speed and efficiency with which they created. The artist was regarded not as a prophet or a genius but as a technically equipped individual who succeeds in beautifying the surroundings of all men. It is in this manner that one can perhaps best define the Muslim artistic tradition: it avoided the conscious search for a unique masterpiece, and it did not build monuments for the eternal glory of God. It sought instead to please man and to make every moment of his life as attractive and enjoyable as possible. There is a hedonistic element in Islāmic art, therefore, but this hedonism is intellectually and emotionally mitigated by the conscious knowledge of the perishable character of all things human. In this fashion, Islāmic art seen as a whole is a curious paradox, for as it softened and embellished life's activities, it was created with destructible materials, thereby reiterating Islām's conviction that only God remains. (O.Gr.)

BIBLIOGRAPHY

Literature: JAMES KRITZECK (comp.), *Modern Islamic Literature: From 1800 to the Present* (1970), is a useful anthology of poetry and prose from different parts of the Muslim world. (*Arabic literature*): CARL BROCKELMANN, *Geschichte der arabischen Litteratur,* 2nd ed. (1943–49, and suppl., 1937–42), is the standard reference work containing information about almost every Arabic writer from pre-Islāmic to modern times. This work has been enlarged by FUAT SEZGİN, *Geschichte des arabischen Schrifttums* (1967–), who has included many hitherto unknown books and manuscripts. R.A. NICHOLSON, *A Literary History of the Arabs,* 2nd ed. (1930, reprinted 1969), emphasizes poetry in the classical age; his *Studies in Islamic Poetry* (1921, reprinted 1963) contains the best analysis of al-Maʿarri's poetry. H.A.R. GIBB, *Arabic Literature,* 2nd rev. ed. (1963), is concise and informative; the German translation, *Arabische Literaturgeschichte* (1968), of Gibb's book has been enlarged by a section on modern Arabic literature by JACOB M. LANDAU and has an extensive bibliography on works of Islāmic literature translated into western European languages. GOTTHOLD WEIL, *Grundriss und System der altarabischen Metren* (1958), is an introduction to Arabic prosody. JOHANN FÜCK, *Arabiya: Untersuchungen zur arabischen Sprach- und Stilgeschichte* (1950), is an indispensable study of the development of a High Arabic style. GUSTAVE E. VON GRUNEBAUM, *Kritik und Dichtkunst: Studien zur arabischen Literaturgeschichte* (1955), contains essays on Arabic literature especially of the ʿAbbāsid period. Von Grunebaum's "Spirit of Islam as Shown in Its Literature," in his *Islam: Essays on the Nature and Growth of a Cultural Tradition,* 2nd ed. (1961, reprinted 1981), is an important essay mainly concerned with Ḥarīrī's *Maqāmāt,* and his "Acculturation as a Theme in Contemporary Arab Literature," in *Diogenes,* 39: 84–118 (1962), is a study on the problem of westernization in modern Arabic literature. ʿABDALQĀHIR AL-JURJĀNĪ, *Die Geheimnisse der Wortkunst* (*Asrār al-balāġa*) . . . (1959), is a German translation by HELLMUT RITTER of this classic on Arabic rhetoric. It is available also in the English translation, *The Mysteries of Eloquence,* ed. by HELLMUT RITTER (1954). ADOLF F. VON SCHACK, *Poesie und Kunst der Araber in Spanien und Sicilien,* 2nd ed., 2 vol. (1877), though often superseded by modern research, remains a charming introduction to the culture and art of Moorish Spain. U.M. DAUDPOTA, *The Influence of Arabic Poetry on the Development of Persian Poetry* (1934), attempts to show the formal influences of Arabic on early Persian poetry. ROGER ALLEN, *The Arabic Novel: An Historical and Critical Introduction* (1982), covers 1938–80. WOLFHART HEINRICHS, *Arabische Dichtung und griechische Poetik* (1969), is an important introduction to the literary criticism of classical Arabic literature.

(*Persian literature*): JAN RYPKA, *History of Iranian Literature* (1968; originally published in Czech, 1956), the standard work on Persian literature from its origins to the 20th century, includes folk literature and Tajik and Indo-Persian literature. See also CHARLES A. STOREY, *Persian Literature: A Bio-Bibliographical Survey,* 2 vol. in 4 (1970–72). EDWARD G. BROWNE, *A Literary History of Persia,* 4 vol. (1902–24, reprinted 1969–78), is an informative classic. A.J. ARBERRY, *Classical Persian Literature* (1958, reprinted 1967), is a classic work by one of the most prolific translators of Arabic and Persian poetry into English. ANTONINO PAGLIARO and ALESSANDRO BAUSANI, *Storia della letteratura Persiana* (1960), contains many interesting and unusual viewpoints. HERMANN ETHÉ, "Neupersische Literatur," and THEODOR NÖLDEKE, "Das iranische Nationalepos," in WILHELM GEIGER and ERNST KUHN (eds.), *Grundriss der iranischen Philologie,* vol. 2 (1896–1904, reprinted 1974), provides a masterly survey of classical Persian literature, including Indo-Persian. FRITZ M. MEIER (ed. and trans.), *Die schöne Mahsatī,* vol. 1 (1963), an immensely learned work centring around the poet Mahsatī, deals with the development of the *rubāʿī* and other forms of Persian poetry. HELLMUT RITTER, *Über die Bildersprache Niẓāmīs* (1927), is the classic work on the imagery

in Neẓāmī's poetry. ANNEMARIE SCHIMMEL, *Stern und Blume* (1984), deals with imagery in Persian poetry. CHARLES H. DE FOUCHÉCOUR, *La Description de la nature dans la poésie lyrique persane du XI^e siècle* (1969), is a study of nature imagery in particular in early Persian poetry. FRIEDRICH RUCKERT, *Grammatik, Poetik und Rhetorik der Perser*, ed. by WILHELM PERTSCH (1874, reprinted 1966), a translation and commentary of a late Indo-Persian manual of rhetoric, is noted for its acute observations and amusing details. FINN THIESEN, *A Manual of Classical Persian Prosody* (1982), is an introduction to problems of Persian, as well as Turkish and Urdu, prosody.

(*Turkish literature*): E.J.W. GIBB, *A History of Ottoman Poetry*, 6 vol. (1900–09, reprinted 1958–67), the classical study of the historical developments of Turkish literature from its beginnings to 1900, includes many translations of poems. OTTO SPIES, *Die Türkische Prosa-literatur der Gegenwart* (1943), deals with Turkish prose after the revolution.

Music: "Bibliography of Asiatic Musics," in the MUSIC LIBRARY ASSOCIATION, *Notes*, 2nd series, vol. 5–6 (1947–49), an extensive bibliography compiled by five scholars, includes an important section on Islāmic music, with 592 references divided into categories dealing with music among Muslims in general, Arabic-speaking peoples, Turkic peoples, and Iranians and others. HENRY GEORGE FARMER, *A History of Arabian Music to the XIIIth Century* (1929, reprinted 1967), is still regarded as a key historical study. His "Music of Islam," in *The New Oxford History of Music*, vol. 1, pp. 421–77 (1957, reprinted 1966), is a good concise survey, as is PETER CROSSLEY-HOLLAND, "The Arabic World," in the *Pelican History of Music*, vol. 1, pp. 118–36 (1960, reprinted 1978). RODOLPHE VON ERLANGER (ed. and trans.), *La Musique Arabe*, 6 vol. (1930–59), includes French translations of the Arabic treatises by al-Fārābī, Avicenna, Ṣafī od-Dīn, and others (vol. 1–4) and devotes the last two volumes to an analytical study of contemporary Arabian music. CURT SACHS, *The Rise of Music in the Ancient World, East and West* (1943), has a large section on Arabic music in the context of an intercultural study. MEHDI BARKECHLI (ed.), *La Musique traditionnelle de l'Iran* (1963), gives a comprehensive musical transcription of the *Radīf* (modal systems of the Iranian traditional music). ADNAN SAYGUN, "La Musique Turque," in the *Encyclopédie de la Pléiade*, vol. 9, pp. 573–617 (1960); and ALEXIS CHOTTIN, *Tableau de la musique marocaine* (1939), discuss regional and local particularities and have useful bibliographies. AMNON SHILOAH, *Caractéristiques de l'art vocal arabe au moyen-âge* (1963), is an important essay on medieval Islāmic vocal music, and his *Theory of Music in Arabic Writings (c. 900–1900)* (1979), is an extensive analytical catalog of manuscripts and published sources on Arabic music. See also O. WRIGHT, *The Model System of Arab and Persian Music, A.D. 1250–1300* (1978), an analytical presentation of the system based on Persian and Arabic medieval treatises; KURT REINHARD and URSULA REINHARD, *Turquie* (1969), a comprehensive presentation of Turkish music in its diverse aspects; and ELLA ZONIS, *Classical Persian Music* (1973), a comprehensive historical study.

Dance and theatre: The classic work on the shadow play in the Middle East is still GEORG JACOB, *Geschichte des Schattentheaters im Morgan- und Abendland*, 2nd ed. (1925, reprinted 1972). METIN AND, *A History of Theatre and Popular Entertainment in Turkey* (1963–64), is a perceptive, scholarly account of the Turkish theatre in all its manifestations, and his *Pictorial History of Turkish Dancing* (1976) is an excellent study on the subject. CHRISTA URSULA SPULER, *Das türkische Drama der Gegenwart* (1968), treats in more detail 20th-century Turkish playwrights and theatrical literature. NICHOLAS N. MARTINOVICH, *The Turkish Theatre* (1933, reprinted 1968); and HELLMUT RITTER, *Karagös*, 3 vol. (1924–53), comprise translations of Turkish shadow plays into English and German, respectively. IGNACZ KUNOS, *Das türkische Volksschauspiel Orta ojnu* (1908), is an introduction to the *ortaoyunu* popular shows, with samples translated into German.

As for the Persian theatre and dance (mainly the latter), the most up-to-date book is MEDJID REZVANI, *Le Théâtre et la danse en Iran* (1962). PETER J. CHELKOWSKI (ed.), *Taʿziyeh: Ritual and Drama in Iran* (1979), is a collection of scholarly writings on the subject. CHARLES VIROLLEAUD, *Le Théâtre persan, ou le drama de Kerbéla* (1950), is a good sampling of *taʿziyah*s in French translation. The Arab theatre and dance (chiefly the former) are discussed in JACOB M. LANDAU, *Studies in the Arab Theater and Cinema* (1958), which also includes a detailed list of Arabic plays.

Visual arts: Among the numerous works dealing with Islāmic art as a whole, only one can be recommended as having a text of considerable merit—KATHARINA OTTO-DORN, *Kunst des Islam* (1964). An important, though partial, interpretation is found in TITUS BURCKHARDT, *Art of Islam: Language and Meaning* (1976). ALEXANDRE PAPADOPOULO, *Islam and Muslim Art* (1979, originally published in French, 1976), presents excellent photographic surveys. See also MILO CLEVELAND BEACH, *The Imperial Image: Paintings from the Mughal Court* (1981). K.A.C. CRESWELL, *A Bibliography of the Architecture, Arts and Crafts of Islam* (1961), with a supplement covering 1960–72 (1973), is a good bibliographical source; current literature on Islāmic art in all languages is surveyed in the *Abstracta Islamica*, published as an annual supplement to the *Revue des Études Islamiques* in Paris. Textual information about the arts has never been properly gathered. For a typical text on painters, see QĀḌĪ AḤMAD, *Calligraphers and Painters*, trans. from the Persian by VLADIMIR MINORSKY (1959). The only lists of artists have been collected by LEO A. MAYER in several books, of which the most important are *Islamic Metalworkers and Their Works* (1959) and *Islamic Architects and Their Works* (1956). The vast majority of material on Islāmic art is to be found in periodicals rather than in books. The three publications that have dealt or deal systematically with all aspects of Islāmic art are *Ars Islamica* (irregular, 1934–51; reprinted in 16 vol., 1968), *Ars Orientalis* (irregular from 1954), and *Kunst des Orients* (annual from 1950). Articles are published in English, French, and German.

Area surveys: (*Spain*): MANUEL GÓMEZ-MORENO, *El arte árabe español hasta los almohades y arte mozárabe* (1951); and LEOPOLDO TORRES BALBÁS, *Arte almohade; arte nazarí; arte mudéjar* (1949). (*North Africa*): There is no recent general work dealing with all the arts; for architecture the indispensable manual is that of GEORGES MARÇAIS, *L'Architecture musulmane d'Occident* (1955), which deals also with Spain. (*Egypt*): DIETRICH BRANDENBURG, *Islamische Baukunst in Ägypten* (1966), is a convenient summary but does not supersede the exhaustive work of K.A.C. CRESWELL, *Muslim Architecture of Egypt*, 2 vol. (1952–59, reprinted 1979), going only up to the middle of the 14th century; and LOUIS HAUTECOEUR and GASTON WIET, *Les Mosquées du Caire*, 2 vol. (1932). A useful periodical is the *Journal of the American Research Center in Egypt*. (*Palestine, Syria*): There are no coherent works dealing with the whole area; JEAN SAUVAGET, *Alep* (1941), is a model (in French) of what can be done with a single city over the centuries; key journals are *Syria* (quarterly), *Levant* (annual), and *Quarterly of the Department of Antiquities in Palestine* (until 1950, when it was superseded in part by the annual publication of the Department of Antiquities of Jordan), and *Annales Archéologiques de Syrie* (annual). (*Iraq and upper Mesopotamia*): The main archaeological source is still FRIEDRICH SARRE and ERNST HERZFELD, *Archäologische Reise im Euphrat- und Tigris-Gebiet*, 4 vol. (1911–20); a model of archaeological history is ROBERT M. ADAMS, *Land Behind Baghdad* (1965). The main journals are *Sumer* (annual) and *Iraq* (semiannual). (*Anatolia*): ESIN ATIL (ed.), *Turkish Art* (1980), covers all fields evenly and has a good bibliography; see also YANNI PETSOPOULOS (ed.), *Tulips, Arabesques and Turbans: Decorative Arts from the Ottoman Empire* (1982). EKREM AKURGAL (ed.), *The Art and Architecture of Turkey* (1980), is a historical treatment of major and minor arts. The principal journals are *Anatolica* (annual) and *Anatolian Studies* (annual). (*Iran*): Nothing has superseded ARTHUR UPHAM POPE and PHYLLIS ACKERMAN (eds.), *A Survey of Persian Art from Prehistoric Times to the Present*, 3rd ed. (1977–). One may also consult ANDRÉ GODARD, *The Art of Iran* (1965, originally published in French, 1962); and the chapters by OLEG GRABAR in *The Cambridge History of Iran*, "The Visual Arts," vol. 4, pp. 329–63 (1975), and "The Visual Arts, 1050–1350," vol. 5, pp. 626–58 (1968). Important information is to be found in HANS E. WULFF, *The Traditional Crafts of Persia* (1966); and in several works by ARTHUR UPHAM POPE, such as *Persian Architecture* (1965). Periodicals of importance are the defunct *Athār-é Īrān* (1936–49), the *Bulletin of the American Institute for Persian Art and Archaeology*, and *Iran* (annual), the active journal of the British School. (*India*): Among several architectural surveys, PERCY BROWN, *Indian Architecture*, vol. 2, *The Islamic Period*, 6th ed. (1971), is the best. See also R. NATH, *History of Sultanate Architecture* (1978), and *History of Mughal Architecture* (1982); WAYNE E. BEGLEY, "Myth of the Taj Mahal and a New Theory of Its Symbolic Meaning," *The Art Bulletin*, 61:7–37 (March 1979); and ELIZABETH B. MOYNIHAN, *Paradise as a Garden* (1979).

Techniques: (*Architecture*): JOHN D. HOAG, *Islamic Architecture* (1976); GEORGE MICHELL (ed.), *Architecture of the Islamic World: Its History and Social Meaning* (1978); and NADER ARDALAN and LALEH BAKHTIAR, *The Sense of Unity: The Sufi Tradition in Persian Architecture* (1973, reprinted 1979). (*Painting*): RICHARD ETTINGHAUSEN, *Arab Painting* (1962); BASIL GRAY, *Persian Painting from Miniatures of the XIII–XVI Centuries* (1947); and DOUGLAS E. BARRETT and BASIL GRAY, *Painting of India* (1963, reissued 1978 as *Indian Painting*). See also OLEG GRABAR, *The Illustrations of the Maqamat* (1984). (*Metalwork*): EVA BAER, *Metalwork in Medieval Islamic Art* (1983); ASSADULLAH SOUREN MELIKIAN-CHIRVANI, *Islamic Metalwork from the Iranian World, 8–18th Centuries* (1982); and

JAMES W. ALLAN, *Islamic Metalwork: The Nuhad Es-Said Collection* (1982). The field owes much to the work of the late D.S. RICE: "Studies in Islamic Metalwork," *Bulletin of the School of Oriental and African Studies, University of London,* vol. 14–16 (1952–57); "Inlaid Brasses from the Workshop of Aḥmad al-Dhakī al-Mawṣilī," *Ars Orientalis,* 2:283–326 (1957); *The Wade Cup in the Cleveland Museum of Art* (1955); and *Le Baptistère de Saint Louis* (1951). See also RICHARD ETTINGHAUSEN, "The Wade Cup . . . ," *Ars Orientalis,* 2:327–366 (1957). (*Ceramics*): The key studies are ARTHUR LANE, *Early Islamic Pottery* (1947, reprinted 1965), and *Later Islamic Pottery,* 2nd ed. (1971). (*Carpets*): Among scholarly studies on carpets are those of KURT ERDMANN, especially *Oriental Carpets* (1960, reissued 1976; originally published in German, 2nd ed., 1960), and *Seven Hundred Years of Oriental Carpets* (1970; originally published in German, 1966). (*Ivories*): JOHN BECKWITH, *Caskets from Cordoba* (1960), is a scholarly study of Moorish ivory work. In addition, see ERNST KÜHNEL, *Die islamischen Elfenbeinskulpturen, VIII.–XIII. Jahrhundert* (1971).

Historical works: (*Early period*): Most of the problems are summarized in OLEG GRABAR, *The Formation of Islamic Art* (1973). For architecture the main books are K.A.C. CRESWELL, *Early Muslim Architecture* (vol. 1, 2nd ed., 1969; vol. 2, 1941; reissued 1979, 2 vol. in 3); R.W. HAMILTON, *Khirbat al-Mafjar* (1959); and JEAN SAUVAGET, *La Mosquée omeyyade de Médine* (1947). (*Middle period*): On the Fāṭimids, see RICHARD ETTINGHAUSEN, "Painting in the Fatimid Period," *Ars Islamica,* 9:112–124 (1942). On the Seljuqs, for Iran, in addition to vol. 4 of *The Cambridge History of Iran,* see RICHARD ETTINGHAUSEN, "Some Comments on Medieval Iranian Art," *Artibus Asiae,* 31:276–300 (1969); for Syria and Egypt, one should consult ERNST HERZFELD, "Damascus," *Ars Islamica,* vol. 9–12 (1942–51); and JEAN SAUVAGET *et al., Les Monuments Ayyoubides de Damas,* 4 vol. (1938–50); and for Anatolia, KURT ERDMANN, *Das anatolische Karavansaray des 13 Jahrhunderts,* 3 vol. (1961–76). Major monuments are discussed by RICHARD ETTINGHAUSEN in "The Bobrinski Kettle," *Gazette des Beaux-Arts,* 24:193–208 (1943); "The Iconography of a Kāshān *Luster* Plate," *Ars Orientalis,* 4:25–64 (1961); "The Flowering of Seljuq Art," *Metropolitan Museum Journal,* 3:113–131 (1970); and ASSADULLAH SOUREN MELIKIAN-CHIRVANI (ed.), *Le Roman de "Varge et Golšâh"* (1970). Newer interpretations of the Alhambra are based on FREDERICK P. BARGEBUHR, *The Alhambra* (1968; originally published in Spanish, 1966); see also OLEG GRABAR, *The Alhambra* (1978). ESIN ATIL, *Renaissance of Islam* (1981), is a comprehensive study of Mamlūk art; see also SALEH L. MOSTAFA, *Kloster und Mausoleum des Farağ ibn Barqūq in Kairo* (1968). For Mongol architecture, see DONALD N. WILBER, *The Architecture of Islamic Iran: The Il Khānid Period* (1955, reprinted 1969); LISA GOLOMBEK, *The Timurid Shrine at Gazur Gah* (1969); and various accounts in the annual *Iran.* For painting, see ERNST J. GRUBE, *The Classical Style in Islamic Painting* (1968), but especially the rich volume of IVAN STCHOUKINE, *Les Peintures des manuscrits timûrides* (1954); and M.S. IPSIROGLU, *Painting and Culture of the Mongols* (1966; originally published in German, 1965). See also OLEG GRABAR and SHEILA BLAIR, *Epic Images and Contemporary History: The Illustrations of the Great Mongol Shahnama* (1980). (*Late period*): For Ottoman architecture, GODFREY GOODWIN, *A History of Ottoman Architecture* (1971); and APTULLAH KURAN, *The Mosque in Early Ottoman Architecture* (1968), supersede all previous work. Painting is covered in NURHAN ATASOY and FILIZ ÇAĞMAN, *Turkish Miniature Painting* (1974). For ceramics, see ARTHUR LANE, "The Ottoman Pottery of Isnik," *Ars Orientalis,* 2:247–282 (1957). For Ṣafavid architecture, see DONALD N. WILBER, *Persian Gardens and Garden Pavilions,* 2nd ed. (1979); RENATA HOLOD (ed.), *Studies on Isfahan* (1974); EUGENIO GALDIERI, *Eṣfahān, 'Ali Qāpū: An Architectural Survey* (1979); MARTIN BERNARD DICKSON and STUART C. WELCH (eds.), *The Houghton Shahnameh* (1981); and ANTHONY WELCH, *Artists for the Shah: Late Sixteenth-Century Painting at the Imperial Court of Iran* (1976). The most important publications on painting are both by IVAN STCHOUKINE, *Les Peintures des manuscrits Safavīs de 1502 à 1587* (1959), and *Les Peintures des manuscrits de Shāh 'Abbās Ier à la fin des Safavīs* (1964). The principal work on India is STUART C. WELCH, *The Art of Mughal India* (1963, reprinted 1976). For contemporary architecture see RENATA HOLOD (ed.), *Architecture and Community* (1983), and the quarterly journal *Mimar,* published in Singapore.

(An.Sc./A.Sh./J.M.L./O.Gr.)

The Islāmic World

Adherence to Islām is a global phenomenon: Muslims predominate in some 30 to 40 countries, from the Atlantic to the Pacific and along a belt that stretches across northern Africa into Central Asia and south to the northern regions of the Indian subcontinent. Arabs account for fewer than one-fifth of all Muslims, more than half of whom live east of Karāchi, Pak. Despite the absence of large-scale Islāmic political entities, the Islāmic faith continues to expand, by some estimates faster than any other major religion.

The Muslim religion and the life of the Prophet Muḥammad are treated specifically in the article ISLĀM, MUḤAMMAD AND THE RELIGION OF. The literature, music, dance, and visual arts of Muslim peoples are treated in the article ISLĀMIC ARTS. Islām is also discussed in articles on individual countries or on regions in which the religion is a factor, such as EGYPT, IRAN, ARABIA, and NORTH AFRICA. Articles on individual branches or sects and concepts are found in the *Micropædia*. See, for example, ISLAM, NATION OF; SUNNITE; SHĪʿITE; ḤADĪTH.

A very broad perspective is required to explain the history of today's Islāmic world. This approach must enlarge upon conventional political or dynastic divisions to draw a comprehensive picture of the stages by which successive Muslim communities, throughout Islām's 14 centuries, encountered and incorporated new peoples so as to produce an international religion and civilization.

In general, events referred to in this article are dated according to the Gregorian calendar, and eras are designated BCE (before the Common Era or Christian Era) and CE (Common Era or Christian Era), terms which are equivalent to BC (before Christ) and AD (Latin: *anno Domini*). In some cases the Muslim reckoning of the Islāmic era is used, indicated by AH (Latin: *anno Hegirae*). The Islāmic era begins with the date of Muḥammad's emigration (*hijrah*) to Medina, which corresponds to July 16, 622, in the Gregorian calendar. The term Islāmic refers to Islām as a religion. The term Islāmicate refers to the social and cultural complex that is historically associated with Islām and the Muslims, even when found among non-Muslims. Islāmdom refers to that complex of societies in which the Muslims and their faith have been prevalent and socially dominant.

The article is divided into the following sections:

Prehistory (c. 3000 BCE–500 CE)

The prehistory of Islāmdom is the history of central Afro-Eurasia from Hammurabi of Babylon to the Achaemenid Cyrus II in Persia to Alexander the Great to the Sāsānian emperor Nūshīrvān to Muḥammad in Arabia; or, in a Muslim view, from Adam to Noah to Abraham to Moses to Jesus to Muḥammad. The potential for Muslim empire building was established with the rise of the earliest civilizations in western Asia. It was refined with the emergence and spread of what have been called the region's Axial Age religions—Abrahamic, centred on the Hebrew patriarch Abraham, and Mazdean, focused on the Iranian deity Ahura Mazdāh—and their later relative, Christianity. It was facilitated by the expansion of trade from eastern Asia to the Mediterranean, and by the political changes thus effected. The Muslims were heirs to the ancient Egyptians, Babylonians, Persians, Hebrews, even the Greeks and Indians; the societies they created bridged time and space, from ancient to modern and from east to west.

THE RISE OF AGRARIAN-BASED CITIED SOCIETIES

In the 7th century CE a coalition of Arab groups, some sedentary and some migratory, inside and outside the Arabian Peninsula, seized political and fiscal control in western Asia, specifically of the lands between the Nile and Oxus (Amu Darya) rivers—territory formerly controlled by the Byzantines in the west and the Sāsānians in the east. The factors that surrounded and directed their accomplishment had begun to coalesce long before, with the emergence of agrarian-based citied societies in western Asia in the 4th millennium BCE. The rise of complex agrarian-based societies, such as Sumer, out of a subsistence agricultural and pastoralist environment, involved the founding of cities, the extension of citied power over surrounding villages, and the interaction of both with pastoralists.

This type of social organization offered new possibilities. Agricultural production and intercity trading, particularly in luxury goods, increased. Some individuals were able to take advantage of the manual labour of others to amass enough wealth to patronize a wide range of arts and crafts; of these, a few were able to establish territorial monarchies and foster religious institutions with wider appeal. Gradually the familiar troika of court, temple, and market emerged. The new ruling groups cultivated skills for administering and integrating non-kin-related groups. They benefited from the increased use of writing and, in many cases, from the adoption of a single writing system, such as the cuneiform, for administrative use. New institutions, such as coinage, territorial deities, royal priesthoods, and standing armies, further enhanced their power.

Court, temple, and market

In such town-and-country complexes the pace of change quickened enough so that a well-placed individual might see the effects of his actions in his own lifetime and be stimulated to self-criticism and moral reflection of an unprecedented sort. The religion of these new social entities reflected and supported the new social environments. Unlike the religions of small groups, the religions of complex societies focused on deities, such as Marduk, Isis, or Mithra, whose appeal was not limited to one small area or group and whose powers were much less fragmented. The relationship of earthly existence to the afterlife became more problematic, as evidenced by the elaborate death rites of Pharaonic Egypt. Individual religious action began to compete with communal worship and ritual; sometimes it promised spiritual transformation and transcendence of a new sort, as illustrated in the pan-Mediterranean mystery religions. Yet large-scale organization had introduced social and economic injustices that rulers and religions could address but not resolve. To many, an absolute ruler uniting a plurality of ethnic, religious, and interest groups offered the best hope of justice.

CULTURAL CORE AREAS OF THE SETTLED WORLD

By the middle of the 1st millennium BCE the settled world had crystallized into four cultural core areas: Mediterranean, Nile-to-Oxus, Indic, and East Asian. The Nile-to-Oxus, the future core of Islāmdom, was the least cohesive and the most complicated. Whereas each of the other regions developed a single language of high culture—Greek, Sanskrit, and Chinese, respectively—the Nile-to-Oxus region was a linguistic palimpsest of Irano-Semitic languages of several sorts: Aramaic, Syriac (eastern or Iranian Aramaic), and Middle Persian (the language of eastern Iran).

The Nile-to-Oxus region. The Nile-to-Oxus region differed in climate and ecology, too. It lay at the centre of a vast arid zone stretching across Afro-Eurasia from the Sahara to the Gobi; it favoured those who could deal with aridity—not only states that could control flooding (as in Egypt), or maintain irrigation (as in Mesopotamia), but also pastoralists and oasis dwellers. Although its agricultural potential was severely limited, its commercial possibilities were virtually unlimited. Located at the crossroads of the trans-Asian trade and blessed with numerous natural transit points, the region offered special social and economic prominence to its merchants.

The period from 800 to 200 BCE has been called the Axial Age because of its pivotal importance for the history of religion and culture. The world's first religions of salvation developed in the four core areas. From these traditions, for example, Judaism, Mazdeism, Buddhism, and Confucianism, derived all later forms of high religion, including Christianity and Islām. Unlike the religions that surrounded their formation, the Axial Age religions concentrated transcendent power into one locus, be it symbolized theistically or nontheistically. Their radically dualistic cosmology posited another realm, totally unlike the earthly realm and capable of challenging and replacing ordinary earthly values. The individual was challenged to adopt the right relationship with that "other" realm, so as to transcend mortality by earning a final resting place, or to escape the immortality guaranteed by rebirth by achieving annihilation of earthly attachment.

The Abrahamic and Mazdean traditions

In the Nile-to-Oxus region two major traditions arose during the Axial Age: the Abrahamic in the west and the Mazdean in the east. Because they required exclusive allegiance through an individual confession of faith in a just and judging deity, they are called confessional religions. The god of these religions was a unique all-powerful creator who remained active in history; and each event in the life of every individual was meaningful in terms of the judgment of God at the end of time. The universally applicable truth of these new religions was expressed in sacred writings. The traditions reflected the mercantile environment in which they were formed in their special concern for fairness, honesty, covenant keeping, moderation, law and order, accountability, and the rights of ordinary human beings. These values were always potentially incompatible with the elitism and absolutism of courtly circles. Most often, as for example in the case of the Achaemenid Empire, the conflict was expressed in rebellion against the crown or was adjudicated by viewing kingship as the guarantor of divine justice.

Although modern Western historiography has projected an East–West dichotomy onto ancient times, Afro-Eurasian continuities and interactions were well established by the Axial Age and persisted throughout premodern times. The history of Islāmdom cannot be understood without reference to them. Through Alexander's conquests in the 4th century BCE in three of the four core areas, the Irano-Semitic cultures of the Nile-to-Oxus region were permanently overlaid with Hellenistic elements, and a link was forged between the Indian subcontinent and Iran. By the 3rd century CE, crosscutting movements like Gnosticism and Manichaeism integrated individuals from disparate cultures. Similarly organized large, land-based empires with official religions existed in all parts of the settled world. The Christian Roman Empire was locked in conflict with its counterpart to the east, the Zoroastrian–Mazdean Sāsānian Empire. Another Christian empire in East Africa, the Abyssinian, was involved alternately with each of the others. In the context of these regional interrelationships inhabitants of Arabia made their fateful entrance into international political, religious, and economic life.

The Arabian Peninsula. The Arabian Peninsula consists of a large central arid zone punctuated by oases, wells, and small seasonal streams and bounded in the south by well-watered lands that are generally thin, sometimes mountainous coastal strips. To the north of the peninsula are the irrigated agricultural areas of Syria and Iraq, the site of large-scale states from the 4th millennium BCE. As early as the beginning of the 1st millennium BCE the southwest corner of Arabia, the Yemen, also was divided into settled kingdoms. Their language was a South Arabian Semitic dialect and their culture bore some affinity to Semitic societies in the Fertile Crescent. By the beginning of the Common Era (the 1st century AD in the Christian calendar) the major occupants of the habitable parts of the arid centre were known as Arabs. They were Semitic-speaking tribes of settled, semi-settled, and fully migratory peoples who drew their name and apparently their identity from what the camel-herding Bedouin pastoralists among them called themselves: *'arab.*

The Arab tribes

Until the beginning of the 3rd century of the Common Era the greatest economic and political power in the peninsula rested in the relatively independent kingdoms of the Yemen. The Yemenis, with a knowledge of the monsoon winds, had evolved an exceptionally long and profitable trade route from East Africa across the Red Sea and from India across the Indian Ocean up through the peninsula into Iraq and Syria, where it joined older Phoenician routes across the Mediterranean and into the Iberian Peninsula. Their power depended on their ability to protect islands discovered in the Indian Ocean and to control the straits of Hormuz and Aden as well as the Bedouin caravanners who guided and protected the caravans that carried the trade northward to Arab entrepôts like Petra and Palmyra. Participation in this trade was in turn an important source of power for tribal Arabs, whose livelihood otherwise depended on a combination of intergroup raiding, agriculture, and animal husbandry.

By the 3rd century, however, external developments began to impinge. In 226 Ardashīr I founded the Sāsānian Empire in Fars; within 70 years the Sāsānian state was at war with Rome, a conflict that was to last up to Islāmic times. The reorganization of the Roman Empire under Constantine the Great, with the adoption of a new faith, Christianity, and a new capital, Constantinople, exacerbated the competition with the Sāsānian Empire and resulted in the spreading of Christianity into Egypt and Abyssinia and the encouraging of missionizing in Arabia itself. There Christians encountered Jews who had been settling since the 1st century, as well as Arabs who had converted to Judaism. By the beginning of the 4th century the rulers of Abyssinia and Ptolemaic Egypt were interfering in the Red Sea area and carrying their aggression into the Yemen proper. In the first quarter of the 6th century the proselytizing efforts of a Jewish Yemeni ruler resulted in a massacre of Christians in the major Christian centre of Najrān. This event invited Abyssinian Christian reprisal and occupation, which put a virtual end to indigenous control of the Yemen. In conflict with the Byzantines, the Zoroastrian–Mazdean Sāsānians invaded Yemen toward the end of the 6th century, further expanding the religious and cultural horizons of Arabia, where membership in a religious community could not be apolitical and could even have international ramifications. The connection between communal affiliation and political orientations would be expressed in the early Muslim community and in fact has continued to function to the present day.

The long-term result of Arabia's entry into international politics was paradoxical: it enhanced the power of the tribal Arabs at the expense of the "superpowers." Living in an ecological environment that favoured tribal independence and small-group loyalties, the Arabs had never established lasting large-scale states, only transient tribal confederations. By the 5th century, however, the settled powers needed their hinterlands enough to foster client states: the Byzantines oversaw the Ghassānid kingdom; the Persians oversaw the Lakhmid; and the Yemenis (prior to the Abyssinian invasion) had Kindah. These relationships increased Arab awareness of other cultures and religions; and the awareness seems to have stimulated internal Arab cultural activity, especially the classical Arabic, or *muḍarī*, poetry, for which the pre-Islāmic Arabs are so famous. In the north, Arabic speakers were drawn into the imperial administrations of the Romans and Sāsānians; soon certain settled and semi-settled Arabs spoke and wrote Aramaic or Persian as well as Arabic, and some Persian or Aramaic speakers could speak and write Arabic. The prosperity of the 5th and 6th centuries, as well as the intensification of imperial rivalries in the late 6th century, seems to have brought the Arabs of the interior permanently into the wider network of communication that fostered the rise of the Muslim community at Mecca and Medina.

Formation and orientation (c. 500–634)

THE CITY OF MECCA: CENTRE OF TRADE AND RELIGION

Although the 6th-century client states were the largest Arab polities of their day, it was not from them that a permanently significant Arab state arose. Rather, it emerged among independent Arabs living in Mecca (Makkah) at the junction of major north–south and west–east routes, in one of the less naturally favoured Arab settlements of the Hejaz (al-Ḥijāz). The development of a trading town into a city-state was not unusual; but unlike many other western Arabian settlements, Mecca was not centred on an oasis or located in the hinterland of any non-Arab power. Although it had enough well water and springwater to provide for large numbers of camels, it did not have enough for agriculture; its economy depended on long-distance as well as short-distance trade.

Mecca under the Quraysh clans. Around the year 400 CE Mecca had come under the control of a group of Arabs who were in the process of becoming sedentary; they were known as Quraysh and were led by a man remembered as Quṣayy. During the generations before Muḥammad's birth in about 570, the several clans of the Quraysh fostered a development in Mecca that seems to have been occurring in a few other Arab towns as well. They used their trading connections and their relationships with their Bedouin cousins to make their town a regional centre whose influence radiated in many directions. They designated Mecca as a quarterly *ḥaram,* a safe haven from the intertribal warfare and raiding that was endemic among the Bedouin. Thus Mecca became an attractive site for large trade fairs that coincided with pilgrimage (*ḥajj*) to a local shrine, the Kaʿbah. The Kaʿbah housed the deities of visitors as well as the Meccans' supra-tribal creator and covenant-guaranteeing deity, called Allāh. Most Arabs probably viewed this deity as one among many, possessing powers not specific to a particular tribe; others may have identified this figure with the God of the Jews and Christians.

Quarterly haven at Mecca

The building activities of the Quraysh threatened one non-Arab power enough to invite direct interference: the Abyssinians are said to have invaded Mecca in the year of Muḥammad's birth. But the Byzantines and Sāsānians were distracted by internal reorganization and renewed conflict; simultaneously the Yemeni kingdoms were declining. Furthermore, these shifts in the international balance of power may have dislocated existing tribal connections enough to make Mecca an attractive new focus for supratribal organization, just as Mecca's equidistance from the major powers protected its independence and neutrality.

The Meccan link between shrine and market has a broader significance in the history of religion. It is reminiscent of changes that had taken place with the emergence of complex societies across the settled world several millennia earlier. Much of the religious life of the tribal Arabs had the characteristics of small-group, or "primitive," religion, including the sacralization of group-specific natural objects and phenomena and the multifarious presence of spirit beings, known among the Arabs as *jinn.* Where more complex settlement patterns had developed, however, widely shared deities had already emerged, such as the "trinity" of Allāh's "daughters" known as al-Lāt, Manāt, and al-ʿUzzāh. Such qualified simplification and inclusivity, wherever they have occurred in human history, seem to have been associated with other fundamental changes—increased settlement, extension and intensification of trade, and the emergence of lingua francas and other cultural commonalties, all of which had been occurring in central Arabia for several centuries.

Changes in religion: the sharing of deities

New social patterns among the Meccans and their neighbours. The sedentation of the Quraysh and their efforts to create an expanding network of cooperative Arabs generated social stresses that demanded new patterns of behaviour. The ability of the Quraysh to solve their problems was affected by an ambiguous relationship between sedentary and migratory Arabs. Tribal Arabs could go in and out of sedentation easily, and kinship ties often transcended life-styles. The sedentation of the Quraysh did not involve the destruction of their ties with the Bedouin or their idealization of Bedouin life. Thus, for example, did wealthy Meccans, thinking Mecca unhealthy, often send their infants to Bedouin foster mothers. Yet the settling of the Quraysh at Mecca was no ordinary instance of sedentation. Their commercial success produced a society unlike that of the Bedouin and unlike that of many other sedentary Arabs. Whereas stratification was minimal among the Bedouin, a hierarchy based on wealth appeared among the Quraysh. Although a Bedouin group might include a small number of outsiders, such as prisoners of war, Meccan society was markedly diverse, including non-Arabs as well as Arabs, slave as well as free. Among the Bedouin, lines of protection for in-group members were clearly drawn; in Mecca, sedentation and socioeconomic stratification had begun to blur family responsibilities and foster the growth of an oligarchy whose economic objectives could easily supersede other motivations and values. Whereas the Bedouin acted in and through groups, and even regularized intergroup raiding and warfare as a way of life, Meccans needed to act in their own interest and to minimize conflict by institutionalizing new, broader social alliances and interrelationships. The market–shrine complex encouraged surrounding tribes to put aside their conflicts periodically and to visit and worship the deities of the Kaʿbah; but such worship, as in most complex societies, could not replace either the particularistic worship of small groups or the competing religious practices of other regional centres, such as aṭ-Ṭāʾif.

Very little in the Arabian environment favoured the formation of stable, large-scale states. Therefore, Meccan efforts at centralization and unification might well have been transient, especially because they were not reinforced by any stronger power and because they depended almost entirely on the prosperity of a trade route that had been formerly controlled at its southern terminus and could be controlled elsewhere in the future, or exclude Mecca entirely. The rise of the Meccan system also coincided with the spread of the confessional religions, through immigration, missionization, conversion, and foreign interference. Alongside members of the confessional religions, unaffiliated monotheists, known as *ḥanīf*s, distanced themselves from the Meccan religious system by repudiating the old gods but embracing neither Judaism nor Christianity.

Concept of social unity through shared deity

Eventually in Mecca and elsewhere a few individuals came to envision the possibility of effecting supra-tribal association through a leadership role common to the confessional religions, that is, prophethood or messengership. The only such individual who succeeded in effecting broad social changes was a member of the Hāshim (Hāshem) clan of Quraysh named Muḥammad ibn ʿAbd Allāh ibn ʿAbd al-Muṭṭalib. One of their own, he accomplished what the Quraysh had started, first by working against them, later by working with them. When he was born, around 570, the potential for pan-Arab unification seemed nil; but after he died, in 632, the first generation of his followers were able not only to maintain pan-Arab unification but to expand far beyond the peninsula.

THE PROPHET MUḤAMMAD

Muḥammad's years in Mecca. *Spiritual awakening.* Any explanation of such an unprecedented development must include an analysis not only of Muḥammad's individual genius but also of his ability to articulate an ideology capable of appealing to multiple constituencies. His approach to the role of prophet allowed a variety of groups to conceptualize and form a single community. Muḥammad was, according to many students of social behaviour, particularly well placed to lead such a social movement; in both ascribed and acquired characteristics he was unusual. Although he was a member of a high-status tribe, he belonged to one of its less well-placed clans. He was fatherless at birth; his mother and grandfather died when he was young, leaving him under the protection of an uncle. Although he possessed certain admirable personality traits to an unusual degree, his commercial success derived not from his own status but from his marriage to a much older woman, a wealthy widow named Khadījah. During the years of his marriage, his personal habits grew increasingly atypical; he began to absent himself in the hills outside Mecca to engage in the solitary spiritual activity of the *ḥanīf*s. At age 40, while on retreat, he saw a figure, whom he later identified as the angel Gabriel, who forced him to repeat these words: "Recite: In the name of God, the Merciful and Compassionate. Recite: And your Lord is Most Generous. He teaches by the pen, teaches man what he knew not." Although a few individuals, including his wife Khadījah, recognized his experience as that of a messenger of God, the contemporary religious life of most of the Meccans and the surrounding Arabs did not prepare them to share in this recognition easily.

Muḥammad's first recitation

Arabs did recognize several other types of intermediaries with the sacred. Some of the kings of the Yemen are said to have had priestly functions; and tribal leaders, *shaykh*s, in protecting their tribes' hallowed custom (*sunnah*), had a spiritual dimension. Tribal Arabs also had their *kāhin*s, religious specialists who delivered oracles in ecstatic rhymed prose (*sajʿ*) and read omens. They also had their *shāʿir*s, professionally trained oral poets who defended the group's honour, expressed its identity, and engaged in verbal duels with the poets of other groups. The power of the recited word was well established; the poets' words were even likened to arrows that could wound the unprotected enemy. Because Muḥammad's utterances seemed similar, at least in form, to those of the *kāhin*s, many of his hearers naturally assumed that he was one of the figures with whom they were more familiar. Indeed, Muḥammad might not even have attracted attention had he not sounded like other holy men; but by eschewing any source other than the one supreme being, whom he identified as Allāh ("the god") and whose message he regarded as cosmically significant and binding, he was gradually able to distinguish himself from all other intermediaries. Like many successful leaders Muḥammad broke through existing restraints by what might be called transformative conservatism. By combining familiar leadership roles with a less familiar one, he expanded his authority; by giving existing practices a new history, he reoriented them; by assigning a new cause to existing problems, he resolved them. His personal characteristics fit his historical circumstances perfectly.

Public recitations. Muḥammad's first vision was followed by a brief lull, after which he began to hear messages frequently, entering a special physical state to receive them and returning to normalcy to deliver them orally. Soon he began publicly to recite warnings of an imminent reckoning by Allāh that disturbed the Meccan leaders. Muḥammad was one of their own, a man respected for his personal qualities. Yet weakening kinship ties and increasing social diversity were helping him attract followers from many different clans and also from among tribeless persons, giving all of them a new and potentially disruptive affiliation. The fundamentals of his message, delivered often in the vicinity of the Kaʿbah itself, questioned the very reasons for which so many people gathered there. If visitors to the Kaʿbah assumed, as so many Arabs did, that the deities represented by its idols were all useful and accessible in that place, Muḥammad spoke, as had Axial Age figures before, of a placeless and timeless deity that not only had created human beings, making them dependent on him, but would also bring them to account at an apocalypse of his own making. In place of time or chance, which the Arabs assumed to govern their destiny, Muḥammad installed a final reward or punishment based on individual actions. Such individual accountability to an unseen power that took no account whatsoever of kin relationships and operated beyond the Meccan system could, if taken seriously, undermine any authority the Quraysh had acquired. Muḥammad's insistence on the protection of the weak, which echoed Bedouin values, threatened the unbridled amassing of wealth so important to the Meccan oligarchy.

Early reactions to Muḥammad's preachings

Efforts to reform Meccan society. Yet Muḥammad also appealed to the town dweller by describing the human being as a member of a polis (city-state) and by suggesting ways to overcome the inequities that such an environment breeds. By insisting that an event of cosmic significance was occurring in Mecca, he made the town the rival of all the greater cities with which the Meccans traded. To Meccans who believed that what went on in their town and at their shrine was hallowed by tribal custom, *sunnah,* Muḥammad replied that their activities in fact were a corrupt form of a practice that had a very long history with the god of whom he spoke. In Muḥammad's view, the Kaʿbah had been dedicated to the aniconic worship of the one God (Allāh) by Abraham, who fathered the ancestor of the Israelites, Isḥāq (Isaac), as well as the ancestor of the Arabs, Ismāʿīl (Ishmael). Muḥammad asked his hearers not to embrace something new, but to abandon the traditional in favour of the original. He appealed to his fellow Quraysh not to reject the *sunnah* of their ancestors, but rather to appreciate and fulfill its true nature. God should be worshiped not through offerings but through prayer and recitation of his messages, and his house should be emptied of its useless idols.

In their initial rejection of his appeal, Muḥammad's Meccan opponents took the first step toward accepting the new idea: they attacked it. For it was their rejection of him, as well as his subsequent rejection by many Jews and Christians, that helped to forge Muḥammad's followers into a community with an identity of its own and capable of ultimately incorporating its opponents. Muḥammad's disparate following was exceptionally vulnerable, bound together not by kinship ties but by a "generic" monotheism that involved being faithful (*muʾmin*) to the message God was sending through their leader. Their vulnerability was mitigated by the absence of formal municipal discipline; but their opponents within Quraysh could apply informal pressures ranging from harassment and violence against the weakest to a boycott against Muḥammad's clan, who were persuaded by his uncle Ṭālib to remain loyal even though most of them were not his followers. Meanwhile Muḥammad and his closest associates were thinking about reconstituting themselves as a separate community in a less hostile environment. In about 612 some 80 of his followers made an emigration (*hijrah*) to Abyssinia, perhaps assuming that they would be welcome in a place that had a history of hostility to the Meccan oligarchy and that worshiped the same god who had sent Muḥammad to them; but they eventually returned without establishing a permanent community. During the next decade, continued rejection intensified the group's identity and its search for another home. Although the boycott against Muḥammad's clan began to disintegrate, the deaths of his wife and his uncle, in about 619, removed an important source of psychological and social support. Muḥammad had already begun to preach and attract followers at market gatherings outside Mecca; now he intensified his search for a more hospitable environment. In 620 he met with a delegation of followers from Yathrib, an oasis about 200 miles to the northeast; in the next two years their support grew into an offer of protection.

Muḥammad's emigration to Yathrib (Medina). Like Mecca, Yathrib was experiencing demographic problems: several tribal groups coexisted, descendants of its Arab Jewish founders as well as a number of pagan Arab immigrants divided into two tribes, the Aws and the Khazraj. Unable to resolve their conflicts, the Yathribis invited Muḥammad to perform the well-established role of neutral outside arbiter (*ḥakam*). In 622, having sent his followers ahead, he and one companion, Abū Bakr, completed the community's second and final emigration, barely avoiding Quraysh attempts to prevent his departure by force. By the time of the emigration a new label had begun to appear in Muḥammad's recitations to describe his followers; in addition to being described in terms of their faithfulness (*īmān*) to God and his messenger, they were also described in terms of their undivided attention, that is, as *muslim*s, individuals who assumed the right relationship to God by surrendering (*islām*) to his will. Although the label *muslim*, derived from *islām*, eventually became a proper name for a specific historical community, at this point it appears to have expressed commonalty with other monotheists: like the others, *muslim*s faced Jerusalem to pray; Muḥammad was believed to have been transported from Jerusalem to the heavens to talk with God; and Abraham, Noah, Moses, David, and Jesus, as well as Muḥammad, all were considered to be prophets (*nabī*s) and messengers of the same God. In Yathrib, however, conflicts between other monotheists and the *muslim*s sharpened their distinctiveness.

Use of the terms *islām* and *muslim*

The forging of Muḥammad's community. As an autonomous community *muslim*s might have become a tribal unit like those with whom they had affiliated, especially because the terms of their immigration gave them no special status. Yet under Muḥammad's leadership they developed a social organization that could absorb or challenge everyone around them. They became Muḥammad's *ummah* ("community") because they had recognized and supported God's emissary (*rasūl Allāh*). The *ummah*'s members differed from one another not by wealth or genealogical superiority but by the degree of their faith and piety; and membership in the community was itself an expression of faith. Anyone could join, regardless of origin, by following Muḥammad's lead, and the nature of members' support could vary. In the concept of *ummah*, Muḥammad supplied the missing ingredient in the Meccan system: a powerful abstract principle for defining, justifying, and stimulating membership in a single community.

Muḥammad made the concept of *ummah* work by expanding his role as arbiter so as to become the sole spokesman for all residents of Yathrib, also known as Medina. Even though the agreement under which Muḥammad had emigrated did not obligate non-Muslims to follow him except in his arbitration, they necessarily became involved in the fortunes of his community. By protecting him from his Meccan enemies, the residents of Medina identified with his fate. Those who supported him as Muslims received special designations: the Medinans were called *anṣār* ("helpers"), and his fellow emigrants were distinguished as *muhājirūn* ("emigrants"). He was often able to use revelation to arbitrate. Because the terms of his emigration did not provide adequate financial support, he began to provide for his community through caravan raiding, a tactic familiar to tribal Arabs. By thus inviting hostility, he required all the Medinans to take sides. Initial failure was followed by success, first at Nakhlah, where the Muslims defied Meccan custom by violating one of the truce months so essential to Meccan prosperity and prestige. Their most memorable victory occurred in 624 at Badr, against a large Meccan force; they continued to succeed, with only one serious setback, at Uḥud in 625. From that time on, "conversion" to Islām involved joining an established polity, the successes of which were tied to its proper spiritual orientation, regardless of whether the convert shared that orientation completely. During the early years in Medina a major motif of Islāmic history emerged: the connection between material success and divine favour, which had also been prominent in the history of the Israelites.

The relation of material success to divine favour

The ummah's allies and enemies. During these years, Muḥammad used his outstanding knowledge of tribal relations to act as a great tribal leader, or *shaykh*, further expanding his authority beyond the role that the Medinans had given him. He developed a network of alliances between his *ummah* and neighbouring tribes, and so competed with the Meccans at their own game. He managed and distributed the booty from raiding, keeping one-fifth for the *ummah*'s overall needs and distributing the rest among its members. In return, members gave a portion of their wealth as *zakat*, to help the needy and to demonstrate their awareness of their dependence on God for all of their material benefits. Like other *shaykh*s, Muḥammad contracted numerous, often strategically motivated, marriage alliances. He was also more able to harass and discipline Medinans, Muslim and non-Muslim alike, who did not support his activities fully; he agitated in particular against the Jews, one of whose clans, the Banū Qaynuqa, he expelled.

Increasingly estranged from nonresponsive Jews and Christians, he reoriented his followers' direction of prayer from Jerusalem to Mecca. He formally instituted the *ḥajj* to Mecca and fasting during the month of Ramaḍān as distinctive cultic acts, in recognition of the fact that *islām*, a generic act of surrender to God, had become Islām, a proper-name identity distinguished not only from paganism but from other forms of monotheism as well. As more and more of Medina was absorbed into the Muslim community, and as the Meccans weakened, Muḥammad's authority expanded. He continued to lead a three-pronged campaign, against nonsupporters in Medina, against the Quraysh in Mecca, and against surrounding tribes; he even ordered raids into southern Syria. Eventually, Muḥammad became powerful enough to punish nonsupporters severely, especially those who leaned toward Mecca. For example, he had the men of the Qurayẓah clan of Jews in Medina executed after they failed to help him against the Meccan forces at the Battle of the Ditch in 627. But he also used force and diplomacy to bring in other Jewish and Christian groups. Because they were seen, unlike pagans, to have formed *ummah*s of their own around a revelation from God, Jews and Christians were entitled to pay for protection (*dhimmah*). Muḥammad thus set a precedent for another major characteristic of Islāmicate civilization, that of qualified religious pluralism under Muslim authority.

Muḥammad's later recitations. During these years of warfare and consolidation, Muḥammad continued to transmit revealed recitations, though their nature began to change. Some commented on Muḥammad's situation, consoled and encouraged his community, explained the continuing resistance of the Meccans, and urged appropriate responses. Some told stories about figures familiar to Jews and Christians, cast in an Islāmic framework. Though still delivered in the form of God's direct speech, the messages became longer and less ecstatic, less urgent in their warnings if more earnest in their guidance. Eventually they focused on interpersonal regulations in areas of particular importance for a new community, such as sexuality, marriage, divorce, and inheritance. By this time certain Muslims had begun to write down what Muḥammad uttered or to recite passages for cultic worship (*ṣalāt*) and private devotion. The recited word, so important among the Arab tribes, had found a greatly enlarged significance. A competitor for Muḥammad's status as God's messenger even declared himself among a nonmember tribe; he was Maslamah of Yamāmah, who claimed to convey revelations from God. He managed to attract numerous Bedouin Arabs but failed to speak as successfully as Muḥammad to the various available constituencies.

The importance of the recited word

Activism in the name of God, both nonmilitary as well as military, would become a permanent strand in Muslim piety. Given the environment in which Muḥammad operated, his *ummah* was unlikely to survive without it; to compete as leader of a community he needed to exhibit military prowess. (Like most successful leaders, however, Muḥammad was a moderate and a compromiser; some of his followers were more militant and aggressive than he, and some were less so.) In addition, circumstantial necessity had ideological ramifications. Because Muḥammad as messenger was also, by divine providence, leader of an established community, he could easily define the whole realm of social action as an expression of faith. Thus Muslims were able to identify messengership with worldly leadership to an extent almost unparalleled in the history of religion. There had been activist prophets before Muḥammad, and there were activist prophets after him, but in no other religious tradition does the image of the activist prophet, and by extension the activist follower, have such a comprehensive and coherent justification in the formative period.

ISLĀM AT MUḤAMMAD'S DEATH

Muḥammad's continuing success gradually impinged on the Quraysh in Mecca. Some defected and joined his community. His marriage to a Quraysh woman provided him with a useful go-between. In 628 he and his followers tried to make an Islāmized *ḥajj* but were forestalled by the Meccans. At al-Ḥudaybiyah, outside Mecca, Muḥammad granted a 10-year truce on the condition that the Meccans would allow a Muslim pilgrimage the next year. Even at this point, however, Muḥammad's control over his followers had its limits; his more zealous followers agreed to the pact only after much persuasion. As in all instances of charismatic leadership, persisting loyalty was correlated with continuing success. In the next year the Meccans allowed a Muslim *ḥajj;* and in the next, 630, the Muslims occupied Mecca without a struggle. Muḥammad began to receive deputations from many parts of Arabia. By his death in 632 he was ruler of virtually all of it.

The Meccan Quraysh were allowed to become Muslims without shame. In fact, they quickly became assimilated to the actual *muhājirūn,* even though they had not emigrated to Yathrib themselves. Ironically, in defeat they had accomplished much more than they would have had they achieved victory: the centralization of all of Arabia around their polity and their shrine, the Kaʿbah, which had been emptied of its idols to be filled with an infinitely greater invisible power.

Because intergroup conflict was banned to all members of the *ummah* on the basis of their shared loyalty to the emissary of a single higher authority, the limitations of the Meccan concept of *ḥaram,* according to which the city quarterly became a safe haven, could be overcome. The broader solidarity that Muḥammad had begun to build was stabilized only after his death; and this was achieved, paradoxically, by some of the same people who had initially opposed him. In the next two years one of his most significant legacies became apparent: the willingness and ability of his closest supporters to sustain the ideal and the reality of one Muslim community under one leader, even in the face of significant opposition. When Muḥammad died, two vital sources of his authority ended—ongoing revelation and his unique ability to exemplify his messages on a daily basis. A leader capable of keeping revelation alive might have had the best chance of inheriting his movement; but no Muslim claimed messengership, nor had Muḥammad unequivocally designated any other type of successor. The *anṣār,* his early supporters in Medina, moved to elect their own leader, leaving the *muhājirūn* to choose theirs; but a small number of *muhājirūn* managed to impose one of their own over the whole. That man was Abū Bakr, one of Muḥammad's earliest followers and the father of his favourite wife, ʿĀʾishah. The title Abū Bakr took, *khalīfah* (caliph), meaning deputy or successor, echoed revealed references to those who assist major leaders and even God himself. To *khalīfah* he appended *rasūl Allāh,* so that his authority was based on his assistance to Muḥammad as messenger of God.

ABŪ BAKR'S SUCCESSION

Abū Bakr soon confronted two new threats: the secession of many of the tribes that had joined the *ummah* after 630 and the appearance among them of other prophet figures who claimed continuing guidance from God. In withdrawing, the tribes appear to have been able to distinguish loyalty to Muḥammad from full acceptance of the uniqueness and permanence of his message. The appearance of other prophets illustrates a general phenomenon in the history of religion: the volatility of revelation as a source of authority. When successfully claimed, it has almost no competitor; once opened, it is difficult to close; and, if it cannot be contained and focused at the appropriate moment, its power disperses. Jews and Christians had responded to this dilemma in their own ways; now it was the turn of the Muslims, whose future was dramatically affected by Abū Bakr's response. He put an end to revelation with a combination of military force and coherent rhetoric. He defined withdrawal from Muḥammad's coalition as ingratitude to or denial of God (the concept of *kufr*); thus he gave secession (*riddah*) cosmic significance as an act of apostasy punishable, according to God's revealed messages to Muḥammad, by death. He declared that the secessionists had become Muslims, and thus servants of God, by joining Muḥammad; they were not free *not* to be Muslims, nor could they be Muslims, and thus loyal to God, under any leader whose legitimacy did not derive from Muḥammad. Finally, he declared Muḥammad to be the last prophet God would send, relying on a reference to Muḥammad in one of the revealed messages as *khatm al-anbiyāʾ* ("Seal of the Prophets"). In his ability to interpret the events of his reign from the perspective of Islām, Abū Bakr demonstrated the power of the new conceptual vocabulary Muḥammad had introduced.

Ban on revelation and secession

Had Abū Bakr not asserted the independence and uniqueness of Islām, the movement he had inherited could have been splintered or absorbed by other monotheistic communities or by new Islām-like movements led by other tribal figures. Moreover, had he not quickly made the ban on secession and intergroup conflict yield material success, his chances for survival would have been very slim, because Arabia's resources could not support his state. To provide an adequate fiscal base, Abū Bakr enlarged impulses present in pre-Islāmic Mecca and in the *ummah.* At his death he was beginning to turn his followers to raiding non-Muslims in the only direction where that was possible, the north. Migration into Syria and Iraq already had a long history; and Arabs, both migratory and settled, were already present there. Indeed some of them were already launching raids when ʿUmar I, Abū Bakr's acknowledged successor, assumed the caliphate in 634. The ability of the Medinan state to absorb random action into a relatively centralized movement of expansion testifies to the strength of the new ideological and administrative patterns inherent in the concept of *ummah.*

The fusion of two once separable phenomena, membership in Muḥammad's community and faith in Islām—the mundane and the spiritual—would become one of Islām's most distinctive features. Becoming and being Muslim always involved *doing* more than it involved *believing.* On balance, Muslims have always favoured orthopraxy (correctness of practice) over orthodoxy (correctness of doctrine). Being Muslim has always meant making a commitment to a set of behavioral patterns because they reflect the right orientation to God. Where choices were later posed, they were posed not in terms of religion and politics, or church and state, but between living in the world the right way or the wrong way. Just as classical Islāmicate languages developed no equivalents for the words *religion* and *politics,* modern European languages have developed no adequate terms to capture the choices as Muslims have posed them.

Fusion of community and faith

Conversion and crystallization (634–870)

SOCIAL AND CULTURAL TRANSFORMATIONS

The Arab conquests are often viewed as a discrete period. The end of the conquests appears to be a convenient dividing line because it coincides with a conventional watershed, the overthrow of the Umayyad caliphs by the ʿAbbāsids. To illustrate their role in broader social and cultural change, however, the military conquests should be included in a period more than twice as long, during which the conquest of the hearts and minds of the majority of the subject population also occurred. Between 634 and 870 Islām was transformed from the badge of a small Arab ruling class to the dominant faith of a vast empire that stretched from the western Mediterranean into Central Asia. As a result of this long and gradual period of conversion, Arab cultures intermingled with the indigenous cultures of the conquered peoples to produce Islām's fundamental orientations and identities. The Arabic language became a vehicle for the transmission of high culture, even though the Arabs remained a minority; for the first time in the history of the Nile-to-Oxus region, a new language of high culture, carrying a great cultural florescence, replaced all previous languages of high culture. Trade and taxation replaced booty as the fiscal basis of the Muslim state; a nontribal army replaced a tribal one; and a centralized empire became a nominal confederation, with all of the social dislocation and rivalries those changes imply.

Arabic as a language of high culture

Yet despite continuous internal dissension, virtually no Muslim raised the possibility of there being more than one legitimate leader. Furthermore, the impulse toward solidarity, inherited from Muḥammad and Abū Bakr, may have actually been encouraged by persisting minority status. While Muslims were a minority, they naturally formed a conception of Islāmic dominance as territorial rather than religious; and of unconverted non-Muslim communities as secondary members. In one important respect the Islāmic faith differed from all other major religious traditions: the formative period of the faith coincided with its political domination of a rich complex of old cultures. As a result, during the formative period of their civilization, the Muslims could both introduce new elements and reorient old ones in creative ways.

Just as Muḥammad fulfilled and redirected ongoing tendencies in Arabia, the builders of early Islāmicate civilization carried forth and transformed developments in the Roman and Sāsānian territories in which they first dominated. While Muḥammad was emerging as a leader in the Hejaz, the Byzantine and Sāsānian emperors were ruling states that resembled what the Islāmicate empire was to become. Byzantine rule stretched from North Africa into Syria and sometimes Iraq; the Sāsānians competed with the Byzantines in Syria and Iraq and extended their sway, at its furthest, across the Oxus River. Among their subjects were speakers and writers of several major languages—various forms of Aramaic such as Mandaean and Syriac; Greek; Arabic; and Middle Persian. In fact, a significant number of persons were probably bilingual or trilingual. Both the Byzantine and Sāsānian empire declared an official religion, Christianity and Zoroastrian–Mazdaism, respectively. The Sāsānian Empire in the early 7th century was ruled by a religion-backed centralized monarchy with an elaborate bureaucratic structure that was reproduced on a smaller scale at the provincial courts of its appointed governors. Its religious demography was complex, encompassing Christians of many persuasions, Monophysites, Nestorians, Orthodox, and others; pagans; gnostics; Jews; Mazdeans. Minority religious communities were becoming more clearly organized and isolated. The population included priests; traders and merchants; landlords (*dihqans*), sometimes living not on the land but as absentees in the cities; pastoralists; and large numbers of peasant agriculturalists. In southern Iraq, especially in and around towns like al-Ḥīrah, it included migratory and settled Arabs as well. Both empires relied on standing armies for their defense and on agriculture, taxation, conquest, and trade for their resources. When the Muslim conquests began, the Byzantines and Sāsānians had been in conflict for a century; in the most recent exchanges, the Sāsānians had established direct rule in al-Ḥīrah, further exposing its many Arabs to their administration. When the Arab conquests began, representatives of Byzantine and Sāsānian rule on Arabia's northern borders were not strong enough to resist.

ʿUMAR I'S SUCCESSION

The spirit of conquest under ʿUmar I. Abū Bakr's successor in Medina, ʿUmar I (ruled 634–644), had not so much to stimulate conquest as to organize and channel it. As leaders he chose skillful managers experienced in trade and commerce as well as warfare and imbued with an ideology that provided their activities with a cosmic significance. The total numbers involved in the initial conquests may have been relatively small, perhaps less than 50,000, divided into numerous shifting groups. Yet few actions took place without any sanction from the Medinan government or one of its appointed commanders. The fighters, or *muqātilah*, could generally accomplish much more with Medina's support than without. ʿUmar, one of Muḥammad's earliest and staunchest supporters, had quickly developed an administrative system of manifestly superior effectiveness. He defined the *ummah* as a continually expansive polity managed by a new ruling elite, which included successful military commanders like Khālid ibn al-Walīd. Even after the conquests ended, this sense of expansiveness continued to be expressed in the way Muslims divided the world into their own zone, the Dār al-Islām, and the zone into which they could and should expand, the Dār al-Ḥarb, the abode of war. The norms of ʿUmar's new elite were supplied by Islām as it was then understood. Taken together, Muḥammad's revelations from God and his *sunnah* (precedent-setting example) defined the cultic and personal practices that distinguished Muslims from others: prayer, fasting, pilgrimage, charity, avoidance of pork and intoxicants, membership in one community centred at Mecca, and activism (*jihād*) in the community's behalf.

ʿUmar I, successor to Abū Bakr

Forging the link of activism with faithfulness. ʿUmar symbolized this conception of the *ummah* in two ways. He assumed an additional title, *amīr al-muʾminīn* ("commander of the faithful"), which linked organized activism with faithfulness (*īmān*), the earliest defining feature of the Muslim. He also adopted a lunar calendar that began with the emigration (*hijrah*), the moment at which a group of individual followers of Muḥammad had become an active social presence. Because booty was the *ummah*'s major resource, ʿUmar concentrated on ways to distribute and sustain it. He established a *dīwān*, or register, to pay all members of the ruling elite and the conquering forces, from Muḥammad's family on down, in order of entry into the *ummah*. The immovable booty was kept for the state. After the government's fifth-share of the movable booty was reserved, the rest was distributed according to the *dīwān*. The *muqātilah* he stationed as an occupying army in garrisons (*amṣār*) constructed in locations strategic to further conquest: al-Fusṭāṭ in Egypt, Damascus in Syria, Kūfah and Basra in Iraq. The garrisons attracted indigenous population and initiated significant demographic changes, such as a population shift from northern to southern Iraq. They also inaugurated the rudiments of an "Islāmic" daily life; each garrison was commanded by a caliphal appointee, responsible for setting aside an area for prayer, a mosque (*masjid*), named for the prostrations (*sujūd*) that had become a characteristic element in the five daily worship sessions (*ṣalāts*). There the fighters could hear God's revelations to Muḥammad recited by men trained in that emerging art. The most pious might commit the whole to memory. There, too, the Friday midday *ṣalāt* could be performed communally, accompanied by an important educational device, the sermon (*khuṭbah*), through which the fighters could be instructed in the principles of the faith. The mosque fused the practical and the spiritual in a special way: because the Friday prayer included an expression of loyalty to the ruler, it could also provide an opportunity to declare rebellion.

Origin of the garrison mosque

The series of ongoing conquests that fueled this system had their most extensive phase under ʿUmar and his successor ʿUthmān ibn ʿAffān (ruled 644–656). Within 25 years, Muslim Arab forces created the first empire permanently to link western Asia with the Mediterranean. Within another century, Muslim conquerors surpassed the achievement of Alexander the Great, not only in the durability of their accomplishment but in its scope as well, reaching from the Iberian Peninsula to Central Asia.

Resistance was generally slight and nondestructive, and conquest through capitulation was preferred to conquest by force. After Sāsānian al-Ḥīrah fell in 633, a large Byzantine force was defeated in Syria, opening the way to the final conquest of Damascus in 636. The next year, further gains were made in Sāsānian territory, especially at the Battle of al-Qādisīyah; in the next, the focus returned to Syria and the taking of Jerusalem. By 640, Roman control in Syria was over; by 641, the Sāsānians had lost all of their territory west of Zagros. During the years 642 to 646 Egypt was taken under the leadership of ʿAmr ibn al-ʿĀṣ, who soon began raids into what the Muslims called the Maghrib, the lands west of Egypt. Shortly thereafter, in the east, Persepolis fell; in 651 the defeat and assassination of the last Sāsānian emperor, Yazdegerd III, marked the end of the 400-year-old Sāsānian Empire.

ʿUTHMĀN'S SUCCESSION AND POLICIES

Discontent in ʿUthmān's reign. This phase of conquest ended under ʿUthmān and ramified widely. ʿUthmān may even have sent an emissary to China in 651; by the end of the 7th century Arab Muslims were trading there. The fiscal strain of such expansion and the growing independence of local Arabs outside the peninsula underlay the persisting discontents that surfaced toward the end of ʿUthmān's reign. The very way in which he was made caliph had already signaled the potential for competition over leadership and resources. Perceived as pliable and docile, he was the choice of the small committee charged by the dying ʿUmar with selecting one of their own number. Once in office, however, ʿUthmān acted to establish the power of Medina over and against some of the powerful Quraysh families at Mecca and local notables outside Arabia. He was accused of nepotism for relying on his own family, the Banū Umayyah, whose talents ʿUmar had already recognized. Among his many other "objectionable" acts was his call for the production of a single standard collection of Muḥammad's messages from God, which was known simply as the Qurʾān ("Recitation" or "Recitations"). Simultaneously he ordered the destruction of any other collections. Although they might have differed only in minor respects, they represented the independence of local communities. Above all, ʿUthmān was the natural target of anyone dissatisfied with the distribution of the conquest's wealth, since he represented and defended a system that defined all income as Medina's to distribute.

ʿUthmān's standardization of the Qurʾān

The difficulties of ʿUthmān's reign took more than a century to resolve. They were the inevitable result not just of the actions of individuals but of the whole process initiated by Muḥammad's achievements. His coalition had been fragile. He had disturbed existing social arrangements without being able to reconstruct and stabilize new ones quickly. Into a society organized along family lines, he had introduced the supremacy of trans-kinship ties. Yet he had been forced to make use of kinship ties himself; and, despite his egalitarian message, he had introduced new inequities by granting privileges to the earliest and most intensely devoted followers of his cause. Furthermore, personal rivalries were stimulated by his charisma; individuals like his wife ʿĀʾishah, his daughter Fāṭimah, and her husband ʿAlī frequently vied for his affection. ʿUmar's *dīwān* had, then, reinforced old inequities by extending privileges to wealthy high-placed Meccans, and it had introduced new tensions by assigning a lower status to those, indigenous or immigrant to the provinces, who joined the cause later (but who felt themselves to be making an equivalent or greater contribution). Other tensions resulted from conditions in the conquered lands: the initial isolation of Arab Muslims, and even Arab Christians who fought with them, from the indigenous non-Arab population; the discouragement of non-Arab converts, except as clients (*mawālī*) of Arab tribes; the administrative dependence of peninsular Arabs on local Arabs and non-Arabs; and the development of a tax system that discriminated against non-Muslims.

Intra-Muslim conflicts. The ensuing conflicts were played out in a series of intra-Muslim disputes that began with ʿUthmān's assassination and continued to the end of the period under discussion. The importance of kinship ties persisted, but they were gradually replaced by the identities of a new social order. These new identities resulted from Muslim responses to anti-Muslim activity as well as from Muslim participation in a series of controversies focused on the issue of leadership. Because the *ummah*, unified under one leader, was seen as an earthly expression of God's favour, and because God was seen as the controller of all aspects of human existence, the identities formed in the course of the *ummah*'s early history could fuse dimensions that secular modern observers are able to distinguish—religious, social, political, and economic. Furthermore, intra-Muslim rivalries changed during the conversion period; the meaningfulness of the new identities expanded as non-Muslims contributed to Islām's formation, through opposition or through conversion, and the key issues broadened as the participating constituencies enlarged. At first the disputes were coterminous with intra-Arab, indeed even intra-Quraysh, rivalries; only later did they involve persons of other backgrounds. Thus the faith of Islām was formed in conjunction with the crises that attended the establishment of rule by Muslims. Muslims might have produced an extremely localized and exclusivistic religion; but in spite of, and perhaps because of, their willingness to engage in continuing internal conflicts, they produced one of the most unified religious traditions in human history.

Unifying of the Muslim tradition

THE FOUR FITNAHS

By the end of the period of conversion and crystallization, Muslim historians would retrospectively identify four discrete periods of conflict and label them *fitnah*s, trials or temptations to test the unity of the *ummah*. Many historians also came to view some identities formed during the *fitnah*s as authentic and others as deviant. This retrospective interpretation may be anachronistic and misleading. The entire period between 656 and the last quarter of the 9th century was conflict-ridden, and the *fitnah*s merely mark periods of intensification; yet the most striking characteristic of the period was the pursuit of unity.

The first fitnah. In the first two *fitnah*s the claimants to the caliphate relied on their high standing among the Quraysh and their local support in either Arabia, Iraq, or Syria. Competition for the caliphate thus reflected rivalries among the leading Arab families as well as regional interests. The first *fitnah* occurred between ʿUthmān's assassination in 656 and the accession of his kinsman Muʿāwiyah I in 661 and included the caliphate of ʿAlī, the cousin and son-in-law of Muḥammad. It involved a three-way contest between ʿAlī's party in Iraq; a coalition of important Quraysh families in Mecca, including Muḥammad's wife ʿĀʾishah and Ṭalḥah and Zubayr; and the party of Muʿāwiyah, the governor of Syria and member of ʿUthmān's clan, the Banū Umayyah. Ostensibly the conflict focused on whether ʿUthmān had been assassinated justly, whether ʿAlī had been involved, and whether ʿUthmān's death should be avenged by Muʿāwiyah or by the leading Meccans. ʿAlī and his party (*shīʿah*) at first gained power over the representatives of the other leading Meccan families, then lost it permanently to Muʿāwiyah, who elevated Damascus, which had been his provincial capital, to the status of imperial capital. Disappointed at the Battle of Ṣiffīn (657) with ʿAlī's failure to insist on his right to rule, a segment of his partisans withdrew, calling themselves accordingly Khawārij (Kharijites; "seceders"). Their spiritual heirs would come to recognize any pious Muslim as leader. Meanwhile, another segment of ʿAlī's party intensified their loyalty to him as a just and heroic leader who was one of Muḥammad's dearest intimates and the father of his only male descendants.

The second fitnah. The second *fitnah* followed Muʿāwiyah's caliphate (661–680), which itself was not free from strife, and coincided with the caliphates of Muʿāwiyah's son Yazīd I (ruled 680–683), whom he designated as successor, and Yazīd's three successors. This *fitnah* was a second-generation reprise of the first; some of the personnel of the former were descendants or relatives of the leaders of the latter. Once again, different regions supported different claimants, as new tribal divisions emerged in the garrison towns; and once again, representatives of the Syrian Umayyads prevailed. In 680, at Karbalā' in Iraq, Yazīd's army murdered al-Ḥusayn, a son of ʿAlī and grandson of Muḥammad, along with a small group of supporters, accusing them of rebellion; and even though the Umayyads subdued Iraq, rebellions in the name of this or that relative of ʿAlī continued, attracting more and more non-Arab support and introducing new dimensions to his cause. In the Hejaz, the Marwānid branch of the Umayyads, descendants of Marwān I who claimed the caliphate in 685, fought against ʿAbd Allāh ibn az-Zubayr for years; by the time they defeated him, they had lost most of Arabia to Kharijite rebels.

Resistance to Muslim authority

During the period of the first two *fitnah*s, resistance to Muslim rule was an added source of conflict. Some of this resistance took the form of syncretic or anti-Islāmic religious movements. For example, during the second *fitnah*, in Iraq a Jew named Abū ʿIsā al-Iṣfahānī led a syncretic movement (that is, a movement combining different forms of belief or practice) on the basis of his claim to be a prophet (an option not generally open to Muslim rebels) and forerunner of the messiah. He viewed Muḥammad, as well as Jesus, as messengers sent not to all humanity but only to their own communities; so he urged each community to continue in its own tradition as he helped prepare for the coming of the messiah. In other areas, such as the newly conquered Maghrib, resistance took the form of large-scale military hostility. In the 660s the Umayyads had expanded their conflict with the Byzantine Empire by competing for bases in coastal North Africa; it soon became clear, however, that only a full-fledged occupation would serve their purposes. That occupation was begun by ʿUqbah ibn Nāfiʿ, the founder of al-Qayrawān (Kairouan, in modern Tunisia) and, as Sidi (Saint) ʿUqbah, the first of many Maghribi Muslim saints. It eventually resulted in the incorporation of large numbers of pagan or Christianized Berber tribes, the first large-scale forcible incorporation of tribal peoples since the secession of tribes under Abū Bakr. But first the Arab armies met fierce resistance from two individuals—one a man, Kusaylah, and one a woman, al-Kāhinah—who became Berber heroes. Berber resistance was not controlled until the end of the 7th century, after which the Berbers participated in the further conquest of the Maghrib and the Iberian peninsula.

During the caliphate of ʿAbd al-Malik ibn Marwān (ruled 685–705), which followed the end of the second *fitnah*, and under his successors during the next four decades, the problematic consequences of the conquests became much more visible. Like their Byzantine and late Sāsānian predecessors, the Marwānid caliphs nominally ruled the various religious communities but allowed the communities' own appointed or elected officials to administer most internal affairs. Yet now the right of religious communities to live in this fashion was justified by the Qur'ān and *sunnah;* as peoples with revealed books (*ahl al-kitāb*), they deserved protection (*dhimmah*) in return for a payment. The Arabs also formed a single religious community whose right to rule over the non-Arab protected communities the Marwānids sought to maintain.

The Dome of the Rock

To signify this supremacy, as well as his co-optation of previous legitimacy, ʿAbd al-Malik ordered the construction of the Dome of the Rock, a monumental mosque, in Jerusalem, a major centre of non-Muslim population. The site chosen was sacred to Jews and Christians because of its associations with biblical history; it held added meaning for Muslims, who believed it to be the starting point for Muḥammad's *miʿrāj* (midnight journey to heaven). Although this and other early mosques resembled contemporary Christian churches, gradually an Islāmic aesthetic emerged: a dome on a geometrical base, accompanied by a minaret from which to deliver the call to prayer; and an emphasis on surface decoration that combined arabesque and geometrical design with calligraphic representations of God's Word. ʿAbd al-Malik took other steps to mark the distinctiveness of Islāmic rule: for example, he encouraged the use of Arabic as the language of government and had Islāmized coins minted to replace the Byzantine and Sāsānian-style coinage that had continued to be used since the conquests. During the Marwānid period, the Muslim community was further consolidated by the regularization of the public cult and the crystallization of a set of five minimal duties (sometimes called pillars).

Yet the Marwānids also depended heavily on the help of non-Arab administrative personnel (*kuttāb;* singular, *kātib*) and on administrative practices (*e.g.*, a set of government bureaus) inherited from Byzantine and, in particular, late Sāsānian practice. Pre-Islāmic writings on governance translated into Arabic, especially from Middle Persian, influenced caliphal style. The governing structure at Damascus and in the provinces began to resemble pre-Islāmic monarchy, and thus appealed to a majority of subjects, whose heritage extolled the absolute authority of a divinely sanctioned ruler. Much of the inspiration for this development came from ʿAbd al-Malik's administrator in the eastern territories, al-Ḥajjāj ibn Yūsuf ath-Thaqafī, who was himself an admirer of Sāsānian practice.

The Marwānid caliphs, as rulers of Muslims and non-Muslims alike, had thus been forced to respond to a variety of expectations. Ironically, it was their defense of the importance and distinctiveness of the Arabic language and the Islāmic community, not their responsiveness to non-Muslim preferences, that prepared the way for the gradual incorporation of most of the subject population into the *ummah*. As the conquests slowed and the isolation of the fighters (*muqātilah*) became less necessary, it became more and more difficult to keep Arabs garrisoned. The sedentation of Arabs that had begun in the Hejaz was being repeated and extended outside the peninsula. As the tribal links that had so dominated Umayyad politics began to break down, the meaningfulness of tying non-Arab converts to Arab tribes as clients was diluted; moreover, the number of non-Muslims who wished to join the *ummah* was already becoming too large for this process to work effectively.

Arabic as a lingua franca

Simultaneously, the growing prestige and elaboration of things Arabic and Islāmic made them more attractive, to non-Arab Muslims and to non-Muslims alike. The more the Muslim rulers succeeded, the more prestige their customs, norms, and habits acquired. Heirs to the considerable agricultural and commercial resources of the Nile-to-Oxus region, they increased its prosperity and widened its horizons by extending its control far to the east and west. Arabic, which occasionally had been used for administrative purposes in earlier empires, now became a valuable lingua franca. As Muslims continued to adapt to rapidly changing circumstances, they needed Arabic to reflect upon and elaborate what they had inherited from the Hejaz. Because the Qur'ān, translation of which was prohibited, was written in a form of Arabic that quickly became archaic to Muslims living in the garrisons, and because it contained references to life in Arabia before and during Muḥammad's time, full understanding of the text required special effort. Scholars began to study the religion and poetry of the *jāhilīyah*, the times of ignorance before God's revelation to Muḥammad. Philologians soon emerged, in the Hejaz as well as in the garrisons. Many Muslims cultivated reports, which came to be known as *ḥadīth*, of what Muḥammad had said and done, in order to develop a clearer and fuller picture of his *sunnah*. These materials were sometimes gathered into accounts of his campaigns, called *maghāzī*. The emulation of Muḥammad's *sunnah* was a major factor in the development of recognizably "Muslim" styles of personal piety and public decision making. As differences in the garrisons needed to be settled according to "Islāmic" principles, the caliphs appointed arbitrating judges, *qāḍī*s, who were knowledgeable in Qur'ān and *sunnah*. The pursuit of legal knowledge, *fiqh*, was taken up in many locales and informed by local pre-Islāmic custom and Islāmic resources. These special forms of knowledge began to be known as *ʿulūm* (singular, *ʿilm*); the persons who pursued them, as *ʿulamā'* (singular, *ʿālim*), a role that provided new sources of prestige and influence, especially for recent converts or sons of converts.

Muslims outside Arabia were also affected by interacting with members of the religious communities over which they ruled. When protected non-Muslims converted, they brought new expectations and habits with them; Islāmic eschatology is one area that reflects such enrichment. Unconverted protected groups (*dhimmī*s) were equally

Interaction with non-Muslims

influential. Expressions of Islāmic identity often had to take into account the critique of non-Muslims, just as the various non-Muslim traditions were affected by contact with Muslims. This interaction had special consequences in the areas of prophethood and revelation, where major shifts and accommodations occurred among Jews, Christians, Mazdeans, and Muslims during the first two centuries of their coexistence. Muslims attempted to establish Muḥammad's legitimacy as an heir to Jewish and Christian prophethood, while non-Muslims tried to distinguish their prophets and scriptures from Muḥammad and the Qur'ān. Within the emergent Islāmicate civilization, the separate religious communities continued to go their own way; but the influence of Muslim rule and the intervention of the caliphs in their internal affairs could not help but affect them. The Babylonian Talmud, completed during these years, bears traces of early interaction among communities. In Iraq caliphal policy helped promote the Jewish gaons (local rabbinic authorities) over the exilarch (a central secular leader). Mazdeans turned to the Nestorian Church to avoid Islām, or reconceptualized Zoroaster as a prophet sent to a community with a Book. With the *dhimmī* system (the system of protecting non-Muslims for payment), Muslim rulers formalized and probably intensified pre-Islāmic tendencies toward religious communalization. Furthermore, the greater formality of the new system could protect the subject communities from each other as well as from the dominant minority. So "converting" to Islām, at least in the Nile-to-Oxus region, meant joining one recognizably distinct social entity and leaving another. One of the most significant aspects of many Muslim societies was the inseparability of "religious" affiliation and group membership, a phenomenon that has translated poorly into the social structures of modern Muslim nations. In the central caliphal lands of the early 8th century, membership in the Muslim community offered the best chance for social and physical mobility, regardless of a certain degree of discrimination against non-Arabs. Among many astounding examples of this mobility is the fact that several of the early governors and independent dynasts of Egypt and the Maghrib were grandsons of men born in Central Asia.

The Marwānid Maghrib illustrates a kind of conversion more like that of the peninsular Arabs. After the defeat of initial Berber resistance movements, the Arab conquerors of the Maghrib quickly incorporated the Berber tribes en masse into the Muslim community, turning them immediately to further conquests. In 710 an Arab–Berber army set out for the Iberian Peninsula under the leadership of Ṭāriq ibn Ziyād (the name Gibraltar is derived from Jabal Ṭāriq, or "Mountain of Ṭāriq"). They defeated King Roderick in 711; raided into and through the Iberian Peninsula, which they called al-Andalūs; and ruled in the name of the Umayyad caliph. The Andalusian Muslims never had serious goals across the Pyrenees. In 732 Charles Martel encountered not a Muslim army but a summer raiding party; despite his "victory" over that party, Muslims continued their seasonal raiding along the southern French coast for many years.

Muslim Andalusia

Muslim Andalusia is particularly interesting because there the pressure for large-scale conversion that was coming to plague the Umayyads in Syria, Iraq, and Iran never developed. Muslims may never have become a majority throughout their 700-year Andalusian presence. Non-Muslims entered into the Muslim realm as Mozarabs, Christians who had adopted the language and manners, rather than the faith, of the Arabs. Given essentially the same administrative arrangements, the Iberian Christian population was later restored to dominance, while the Syrian Christian population was drastically reduced; but the Iberian Jewish population all but disappeared while the Nile-to-Oxus Jewish population survived.

The Berbers who remained in the Maghrib illustrate the mobility of ideologies and institutions from the central lands to more recently conquered territories. No sooner had they given up anti-Muslim resistance and joined the Muslim community than they rebelled again; but this time an Islāmic identity, Kharijism, provided the justification. Kharijite ideas had been carried to the Maghrib by refugees from the numerous revolts against the Marwānids. Kharijite egalitarianism suited the economic and social grievances of the Berbers as non-Arab Muslims under Arab rule. The revolts outlasted the Marwānids; they resulted in the first independent Maghribi dynasty, the Rustamid, founded by Muslims of Persian descent. The direct influence of the revolts was felt as late as the 10th century and survives among small communities in Tunisia and Algeria.

The third fitnah. Meanwhile, in the central caliphal lands, growing discontent with the emerging order crystallized in a multifaceted movement of opposition to the Marwānids. It culminated in the third *fitnah* (744–750), which resulted in the establishment of a new and final dynasty of caliphs, the 'Abbāsids. Ever since the second *fitnah,* a number of concerned and self-conscious Muslims had begun to raise serious questions about the proper Muslim life and the Marwānids' ability to exemplify it, and to answer them by reference to key events in the *ummah*'s history. Pious Muslims tried to define a good Muslim and to decide whether a bad Muslim should be excluded from the community, or a bad caliph from office. They also considered God's role in determining a person's sinfulness and final dispensation. The proper relationship between Arab and non-Arab Muslims, and between Muslims and *dhimmī*s, was another important and predictable focus of reflection. The willingness of non-Arabs to join the *ummah* was growing; but the Marwānids had not found a solution that was either ideologically acceptable or fiscally sound. Because protected non-Muslim groups paid special taxes, fiscal stability seemed to depend on continuing to discourage conversion. One Marwānid, 'Umar II (ruled 717–720), experimented unsuccessfully with a just solution. In these very practical and often pressing debates lay the germs of Muslim theology, as various overlapping positions, not always coterminous with political groupings, were taken: rejecting the history of the community by demanding rule by Muḥammad's family; rejecting the history of the community by following any pious Muslim and excluding any sinner; or accepting the history of the community, its leaders, and most of its members.

In the course of these debates the Marwānid caliphs began to seem severely deficient to a significant number of Muslims of differing persuasions and aspirations. Direct and implied criticism began to surface. Al-Ḥasan al-Baṣrī, a pious ascetic and a model for the early *ṣūfī*s, called on the Marwānids to rule as good Muslims, and on good Muslims to be suspicious of worldly power. Ibn Isḥāq composed an account of Muḥammad's messengership that emphasized the importance of the *anṣār,* the Yathribi tribes that accepted Muḥammad, and by implication the non-Arab converts (from whom Ibn Isḥāq himself was descended).

Criticism of the Marwānid caliphs

The Marwānids were accused of *bid'ah,* new actions for which there were no legitimate Islāmic precedents. Their continuation of pre-Islāmic institutions—the spy system, extortion of deposed officials by torture, and summary execution—were some of their most visible "offenses." To the pious, the ideal ruler, or *imām* (the word also for a Muslim who led the *ṣalāt*), should, like Muḥammad, possess special learning and knowledge. The first four caliphs, they argued, had been *imām*s in this sense; but under the Umayyads the caliphate had been reduced to a military and administrative office devoid of *imāmah,* of true legitimacy. This piety-minded opposition to the Umayyads, as it has been aptly dubbed, now began to talk about a new dispensation. Some of the most vocal members found special learning and knowledge only in Muḥammad's family. Some defined Muḥammad's family broadly to include any Hāshimite; others, more narrowly, to include only descendants of 'Alī. As the number of Muḥammad's descendants through 'Alī had grown, numerous rebellions had broken out in the name of one or the other, drawing on various combinations of constituencies and reflecting a wide spectrum of Islāmic and pre-Islāmic aspirations.

In the late Marwānid period, the piety-minded opposition found expression in a movement organized in Khorāsān (Khurasan) by Abū Muslim, a semisecret operative of one particularly ambitious Hāshimite family, the 'Abbāsids. The 'Abbāsids, who were kin but not descendants of

Muhammad, claimed also to have inherited, a generation earlier, the authority of one of ʿAlī's actual descendants, Abū Hāshim. Publicly Abū Muslim called for any qualified member of Muhammad's family to become caliph; but privately he allowed the partisans (*shīʿah*) of ʿAlī to assume that he meant them. Abū Muslim ultimately succeeded because he managed to link the concerns of the piety-minded in Syria and Iraq with Khorāsānian discontent. He played upon the grievances of its Arab tribes against the tribes of Syria and their representatives in the Khorāsānian provincial government, and on the millennial expectations of non-Arab converts and non-Muslims disenchanted with the injustices of Marwānid rule.

When in 750 the army organized and led by Abū Muslim succeeded in defeating the last Marwānid ruler, his caliph-designate represented only one segment of this broad coalition. He was the head of the ʿAbbāsid family, Abū al-ʿAbbās as-Saffāh, who now subordinated the claims of the party of ʿAlī to those of his own family, and who promised to restore the unity of the *ummah*, or *jamāʿah*. The circumstances of his accession reconfigured the piety-minded opposition that had helped bring him to power. The party (*shīʿah*) of ʿAlī refused to accept the compromise the ʿAbbāsids offered. Their former fellow-opponents did accept membership in the reunified *jamāʿah*, isolating the People of the Shīʿah and causing them to define themselves in terms of more radical points of view. Those who accepted the early ʿAbbāsids came to be known as the People of the Sunnah and Jamāʿah. They accepted the cumulative historical reality of the *ummah*'s first century: all of the decisions of the community, and all of the caliphs it had accepted, had been legitimate, as would be any subsequent caliph who could unite the community. The concept of *fitnah* acquired a fully historicist meaning: if internal discord were a trial sent by God, then any unifying victor must be God's choice.

Sunnites and Shīʿites. The historicists came to be known as Sunnites, their main opponents, as Shīʿites. These labels are somewhat misleading, because they imply that only the Sunnites tried to follow the *sunnah* of Muhammad. In fact, each group relied on the *sunnah*, but emphasized different elements. For the Sunnites, who should more properly be called the Jamāʿi-Sunnites, the principle of solidarity was essential to the *sunnah*. The Shīʿites argued that the fundamental element of the *sunnah*, and one willfully overlooked by the Jamāʿi-Sunnites, was Muhammad's devotion to his family and his wish that they succeed him through ʿAlī. These new labels expressed and consolidated the social reorganization that had been under way since the beginning of the conquests. The vast majority of Muslims now became consensus-oriented, while a small minority became oppositional. The inherent inimitability of Muhammad's role had made it impossible for any form of successorship to capture universal approval.

Crystallization of Shīʿite movement

When the ʿAbbāsids denied the special claims of the family of ʿAlī, they prompted the Shīʿites to define themselves as a permanent opposition to the status quo. The crystallization of Shīʿism into a movement of protest received its greatest impetus during and just after the lifetime of one of the most influential Shīʿite leaders of the early ʿAbbāsid period, Jaʿfar ibn Muhammad (also called Jaʿfar aṣ-Ṣādiq; 765). Jaʿfar's vision and leadership allowed the Shīʿites to understand their chaotic history as a meaningful series of efforts by truly pious and suffering Muslims to right the wrongs of the majority. The leaders of the minority had occupied the office of *imām*, the central Shīʿite institution, which had been passed on from the first *imām*, ʿAlī, by designation down to Jaʿfar, the sixth. To protect his followers from increasing Sunnite hostility to the views of radical Shīʿites, known as the *ghulāt* ("extremists"), who claimed prophethood for ʿAlī, Jaʿfar made a distinction that both protected the uniqueness of prophethood and established the superiority of the role of *imām*. Since prophethood had ended, its true intent would die without the *imāms*, whose protection from error allowed them to carry out their indispensable task.

Although Jaʿfar did develop an ideology that invited Sunnite toleration, he did not unify all Shīʿites. Differences continued to be expressed through loyalty to various of his relatives. During Jaʿfar's lifetime, his uncle Zayd revolted in Kūfah (740), founding the branch of the Shīʿism known as the *zaydīyah* (Zaydis), or Fivers (for their allegiance to the fifth *imām*), who became particularly important in southern Arabia. Any pious follower of ʿAlī could become their *imām*, and any *imām* could be deposed if he behaved unacceptably. The Shīʿite majority followed Jaʿfar's son Mūsā al-Kāẓim and *imāms* in his line through the 12th, who disappeared in 873. Those loyal to the 12 *imāms* became known as the Imāmīs or Ithnā ʿAsharīyah (Twelvers). They adopted a quietistic stance toward the status quo government of the ʿAbbāsids and prepared to wait until the 12th *imām* should return as the messiah to avenge injustices against Shīʿites and to restore justice before the Last Judgment. Some of Jaʿfar's followers, however, remained loyal to Ismāʿīl, Jaʿfar's eldest son who predeceased his father after being designated. These became the Ismāʿīlīyah (Ismāʿīlis) or Sabʿīyah (Seveners), and they soon became a source of continuing revolution in the name of Ismāʿīl's son Muhammad at-Tamm, who was believed to have disappeared. Challenges to the ʿAbbāsids were not long in coming; of particular significance was the establishment, in 789, of the first independent Shīʿite dynasty, in present-day Morocco, by Idrīs ibn ʿAbd Allāh ibn Hasan II, who had fled after participating in an unsuccessful uprising near Mecca. Furthermore, Kharijite rebellions continued to occur regularly.

Branches of Shīʿism

Legitimacy was a scarce and fragile resource in all premodern societies; in the early ʿAbbāsid environment, competition to define and secure legitimacy was especially intense. The ʿAbbāsids came to power vulnerable; their early actions undermined the unitive potential of their office. Having alienated the Shīʿites, they liquidated the Umayyad family, one of whom, ʿAbd ar-Rahmān I, escaped and founded his own state in Andalusia. Although the ʿAbbāsids were able to buttress their legitimacy by employing the force of their Khorāsānian army, by appealing to their piety-minded support, and by emphasizing their position as heirs to the pre-Islāmic traditions of rulership, their own circumstances and policies militated against them. Despite their continuing preference for Khorāsānian troops, the ʿAbbāsids' move to Iraq and their execution of Abū Muslim disappointed the Khorāsānian chauvinists who had helped them. The non-Muslim majority often rebelled, too. Bih'āfrīd ibn Farwardīn claimed to be a prophet capable of incorporating both Mazdaism and Islām into a new faith. Hāshim ibn Hākim, called al-Muqannaʿ (the "Veiled One"), around 759 declared himself a prophet and then a god, heir to all previous prophets, to numerous followers of ʿAlī, and to Abū Muslim himself.

The ʿAbbāsid court at Baghdad

The ʿAbbāsids symbolized their connection with their pre-Islāmic predecessors by founding a new capital, Baghdad, near the old Sāsānian capital. They also continued to elaborate the Sāsānian-like structure begun by the Marwānid governors in Iraq. Their court life became more and more elaborate, the bureaucracy fuller, the inner sanctum of the palace fuller than ever with slaves and concubines as well as the retinues of the caliph's four legal wives. By the time of Hārūn ar-Rashīd (ruled 786–809), Europe had nothing to compare with Baghdad, not even the court of his contemporary Charlemagne (742–814). But problems surfaced, too. Slaves' sons fathered by Muslims were not slaves and so could compete for the succession. Despite the ʿAbbāsids' defense of Islām, unconverted Jews and Christians could be influential at court. The head (vizier or *wazīr*) of the financial bureaucracy sometimes became the effective head of government by taking over the chancery as well. Like all absolute rulers, the ʿAbbāsid caliphs soon confronted the insoluble dilemma of absolutism: the monarch cannot be absolute unless he depends on helpers, but his dependence on helpers undermines his absolutism. Hārūn ar-Rashīd experienced this paradox in a particularly painful way: having drawn into his service prominent members of a family of Buddhist converts, the Barmakids, he found them such rivals that he liquidated them within a matter of years. It was also during Hārūn's reign that Ibrāhīm ibn al-Aghlab, a trusted governor in Tunis, founded a dynasty that gradually became independent, as did the Ṭāhirids, the ʿAbbāsid governors in Khorāsān, two decades later.

The ʿAbbāsids' ability to rival their pre-Islāmic predecessors was enhanced by their generous patronage of artists and artisans of all kinds. The great 7,000-mile Silk Road from Ch'ang-an (now Sian, China) to Baghdad (then the two largest cities in the world) helped provide the wealth. The ensuing literary florescence was promoted by the capture of a group of Chinese papermakers at the Battle of Talas in 751. The ʿAbbāsids encouraged translation from pre-Islāmic languages, particularly Middle Persian, Greek, and Syriac. This activity provided a channel through which older thought could enter and be reoriented by Islāmicate societies. In the field of mathematics, al-Khwārizmī, from whose name the word *algorithm* is derived, creatively combined Hellenistic and Sanskritic concepts. The word *algebra* derives from the title of his major work, *Kitab al-jabr wa al-muqābalah* ("The Book of Integration and Equation"). Movements such as *falsafah* (a combination of the positive sciences with logic and metaphysics) and *kalām* (systematic theological discourse) applied Hellenistic thought to new questions. The translation of Indo-Persian lore promoted the development of *adab,* a name for a sophisticated prose literature as well as the set of refined urbane manners that characterized its clientele. Soon a movement called *shuʿūbīyah* arose to champion the superiority of non-Arabic tastes over the alleged crudeness of the poetry so dear to Arabic litterateurs. However, the great writer of early ʿAbbāsid times, al-Jāḥiẓ, produced a type of *adab* that fused pre-Islāmic and Islāmic concerns in excellent Arabic style. Many of these extra-Islāmic resources conflicted with Islāmic expectations. Ibn al-Muqaffaʿ, an administrator under al-Manṣūr (ruled 754–775), urged his master to emulate pre-Islāmic models, lest the law that the religious specialists (the *ʿulamāʾ*) were developing undermine caliphal authority irrevocably.

The ʿAbbāsids never acted on such advice completely; they even contravened it by appealing for piety-minded support. Having encouraged conversion, they tried to "purify" the Muslim community of what they perceived to be socially dangerous and alien ideas. Al-Mahdī (ruled 775–785) actively persecuted the Manichaeans, whom he defined as heretics so as to deny them status as a protected community. He also tried to identify Manichaeans who had joined the Muslim community without abandoning their previous ideas and practices. ʿAbbāsid "purification of Islām" ironically coincided with some of the most significant absorption of pre-Islāmic monotheistic lore to date, as illustrated by the stories of the prophets written by Al-Kisaʾi, grammarian and tutor to a royal prince. Even though, like the Marwānids, the ʿAbbāsids continued to maintain administrative courts, not accessible to the *qāḍī*s, they also promoted the study of *ʿilm* and the status of those who pursued it. In so doing they fostered what Ibn al-Muqaffaʿ had feared—the emergence of an independent body of law, Sharīʿah, which Muslims could use to evaluate and circumvent caliphal rule itself.

Sharīʿah

A key figure in the development of Sharīʿah was Abū ʿAbd Allāh ash-Shāfiʿī, who died in 820. By his time Islāmic law was extensive but uncoordinated, reflecting differing local needs and tastes. Schools had begun to form around various recognized masters, such as al-Awzāʿī in Syria, Abū Ḥanīfah in Iraq, and Mālik ibn Anas, all of whom used some combination of local custom, personal reasoning, Qurʾān, and Ḥadīth. Ash-Shāfiʿī was born in Mecca, studied with Mālik, participated in a Shīʿite revolt in the Yemen, and was sent to Baghdad as a prisoner of the caliph. After his release he emigrated to Egypt, where he produced his most famous work. Like most other *faqih*s (students of jurisprudence, or *fiqh*), ash-Shāfiʿī viewed Muḥammad's community as a social ideal and his first four successors as rightly guided. So that this exemplary time could provide the basis for Islāmic law, he constructed a hierarchy of legal sources: Qurʾān; Ḥadīth, clearly traceable to Muḥammad and in some cases to his companions; *ijmāʿ* (consensus); and *qiyās* (analogy to one of the first three).

The way in which Islāmic law had developed had allowed many pre-Islāmic customs, such as the veiling and seclusion of women, to receive a sanction not given to them in the Qurʾān or the Ḥadīth. Ash-Shāfiʿī did not change that entirely. Law continued to be pursued in different centres, and several major "ways" (*madhhabs*) began to coalesce among Sunnites and Shīʿites alike. Among Sunnites, four schools came to be preeminent, Shāfiʿīyah (Shafiites), Mālikīyah (Malikites), Ḥanafīyah (Hanafites), and Ḥanābilah (Hanbalites), and each individual Muslim was expected to restrict himself to only one. Furthermore, the notion that the gate of *ijtihād* (personal effort at reasoning) closed in the 9th century was not firmly established until the 12th century. However, ash-Shāfiʿī's system was widely influential in controlling divergence and in limiting undisciplined forms of personal reasoning. It also stimulated the collecting and testing of *ḥadīth* for their unbroken traceability to Muḥammad or a companion. The need to verify *ḥadīth* stimulated a characteristic form of premodern Muslim intellectual and literary activity, the collecting of biographical materials into compendiums (*ṭabaqāt*). By viewing the Qurʾān and documentable *sunnah* as preeminent, ash-Shāfiʿī also undermined those in ʿAbbāsid court circles who wanted a more flexible base from which the caliph could operate. The Sharīʿah came to be a supremely authoritative, comprehensive set of norms and rules covering every aspect of life, from worship to personal hygiene. It applied equally to all Muslims, including the ruler, whom Sharīʿah-minded Muslims came to view as its protector, not its administrator or developer. While the caliphs were toying with theocratic notions of themselves as the shadow of God on earth, the students of legal knowledge were defining their rule as "nomocratic," based only on the law they protected and enforced.

According to the Sharīʿah, a Muslim order was one in which the ruler was Muslim and the Sharīʿah was enshrined as a potential guide to all; Muslims were one confessional community among many, each of which would have its own laws that would apply except in disputes between members of different communities. The Sharīʿah regulated relations and inequities among different segments of society, freeborn Muslim, slave, and protected non-Muslim. The process that produced Sharīʿah resembled the evolution of Oral Torah and rabbinic law, which the Sharīʿah resembled in its comprehensiveness, egalitarianism, and consensualism, in its absorption of local custom, in its resistance to distinguishing the sublime from the mundane, and in its independence from government. Like many Jews, many ultra-pious Muslims came to view the law as a divine rather than human creation.

The fourth fitnah. During the reign of al-Maʾmūn (813–833) the implications of all this *ʿilm*-based activity for caliphal authority began to become clear. Al-Maʾmūn came to the caliphate as the result of the fourth *fitnah,* which reflected the persisting alienation of Khorāsān. Al-Maʾmūn's father, Hārūn ar-Rashīd, provided for the empire to be divided at his death between two sons. Al-Amīn would rule in the capital and all the western domains; al-Maʾmūn, from his provincial seat at Merv in Khorāsān, would rule the less significant east. When Hārūn died, his sons struggled to expand their control; al-Maʾmūn won. During his reign, which probably represents the high point of caliphal absolutism, the court intervened in an unprecedented manner in the intellectual life of its Muslim subjects, who for the next generation engaged in the first major intra-Muslim conflict that focused on belief as well as practice. The Muslims, who now constituted a much more sizable proportion of the population but whose faith lacked doctrinal clarity, began to engage in an argument reminiscent of 2nd-century Christian discussions of the *logos.* Among Christians, for whom the Word was Jesus, the argument had taken a Christological form. But for Muslims the argument had to centre on the Qurʾān and its created or uncreated nature. Al-Maʾmūn, as well as his brother and successor al-Muʿtaṣim, was attracted to the Muʿtazilah (Mutazilites), whose school had been influenced by Hellenistic ideas as well as by contact with non-Muslim theologians. If the Qurʾān were eternal along with God, his unity would, for the Muʿtazilah, be violated. They especially sought to avoid literal exegesis of the Qurʾān, which in their view discouraged free will and

Debate on the nature of the Qurʾān

produced embarrassing inconsistencies and anthropomorphisms. By arguing that the Qur'ān was created in time, they could justify metaphorical and changing interpretation. By implication, Muḥammad's position as deliverer of revelation was undermined because *ḥadīth* was made less authoritative.

The opponents of the Muʿtazilah, and therefore of the official position, coalesced around the figure of Aḥmad ibn Ḥanbal. A leading master of *ḥadīth*, he had many followers, some of them recent converts, whom he was able to mobilize in large public demonstrations against the doctrine of the created Qur'ān. Because viewing the Qur'ān as created would invalidate its absolute authority, Ibn Ḥanbal argued for an eternal Qur'ān and emphasized the importance of Muḥammad's *sunnah* to the understanding of it. By his time, major literary works had established a coherent image of the indispensability of Muḥammad's prophethood; in fact, just before the Muʿtazilite controversy began, Ibn Hishām had produced his classic recension of the *Sīrah*, or life, of Muḥammad, composed half a century earlier by Ibn Isḥāq. As in the early Christian Church, these were not merely dogmatic issues. They were rooted in the way ordinary Muslims lived, just as affection for a divine Christ had become popular sentiment by the time Arius and Athanasius debated. Although Muslims lacked an equivalent of the Christian Church, they resolved these issues similarly; like Jesus for the Christians, the Qur'ān for the Muslims was somehow part of God; *ḥadīth*-mindedness and emulation of Muḥammad's *sunnah* had become such an essential part of the daily life of ordinary people that the Muʿtazilite position, as intellectually consistent and attractive as it was, was unmarketable. In a series of forcible inquiries called *miḥnah*, al-Ma'mūn and al-Muʿtaṣim actively persecuted those who, like Ibn Ḥanbal, would not conform; but popular sentiment triumphed and after al-Muʿtaṣim's death the caliph al-Mutawakkil was forced to reverse the stand of his predecessors.

This caliphal failure to achieve doctrinal unity coincided with other crises. By al-Muʿtaṣim's reign the tribal troops were becoming unreliable and the Ṭāhirid governors of Khorāsān more independent. Al-Muʿtaṣim expanded his use of military slaves, finding them more loyal but more unruly, too. Soon he had to house them at Sāmarrā', a new capital north of Baghdad, where the caliphate remained until 892. For most of this period, the caliphs were actually under the control of their slave soldiery; and even though they periodically reasserted their authority, rebellions continued. Many were anti-Muslim, like that of the Iranian Bābak (whose 20-year-long revolt was crushed in 837); but increasingly they were intra-Muslim, like the Kharijite-led revolt of black agricultural slaves (Zanj) in southern Iraq (868–883). By 870, then, the Baghdad–Sāmarrā' caliphate had become one polity among many; its real rulers had no ideological legitimacy. At Córdoba the Umayyads had declared their independence; and the Maghrib was divided among several dynasties of differing persuasions, the Shīʿite Idrīsids, the Kharijite Rustamids, and the Jamāʿi-Sunnite Aghlabids. The former governors of the ʿAbbāsids, the Ṭūlūnids, ruled Egypt and parts of Arabia; Iran was divided between the Ṣaffārids, governors of the ʿAbbāsids in the south, and the Persian Sāmānids in the north.

Unifying forces in Islāmdom

The centrifugal forces represented by these administrative divisions should not obscure, however, the existence of numerous centripetal forces that continued to give Islāmdom, from Andalusia to Central Asia, other types of unity. The ideal of the caliphate continued to be a source of unity after the reality waned; among all the new states, no alternative to the caliphate could replace it. Furthermore, now that Muslims constituted a majority almost everywhere in Islāmdom, conflict began to be expressed almost exclusively in Islāmic rather than anti-Islāmic forms. In spite of continuing intra-Muslim conflict, Muslim worship and belief remained remarkably uniform. The annual pilgrimage to Mecca helped reinforce this underlying unity by bringing disparate Muslims together in a common rite. The pilgrimage, as well as the rise of prosperous regional urban centres, enhanced the trade that traversed Islāmdom regardless of political conflicts; along the trade routes that crisscrossed Eurasia, Islāmdom at its centre, moved not only techniques and goods but ideas as well. A network of credit and banking, caravansaries, and intercity mercantile alliances, tied far-flung regions together. Central was the caravan, then the world's most effective form of transport. The peripatetic nature of education promoted cross-fertilization. Already the *faqīr* (fakir), a wandering mendicant Ṣūfī dervish, was a familiar traveler. Across Islāmdom, similar mosque–market complexes sprang up in most towns; because municipal institutions were rare, political stability so unpredictable, and government intervention kept to a minimum (sometimes by design, more often by necessity), the Sharīʿah and the learned men who carried it became a mainstay of everyday life and social intercourse. The Sharīʿah, along with the widespread affection for the *sunnah* of Muḥammad, regulated, at least among pious Muslims, personal habits of the most specific sort, from the use of scent to the cut of a beard. Comprehensive and practical, the *sunnah* could amuse as well. When asked whether to trust in God or tie one's camel, so a popular *ḥadīth* goes, the Prophet replied, "Trust in God, then tie your camel."

The significance of *ḥadīth* and *sunnah* is represented by the ending date of the period of conversion and crystallization. No one can say exactly when the majority of Islāmdom's population became Muslim. Older scholarship looks to the end of the first quarter of the 9th century; newer scholarship to the beginning of the third quarter. In 870 a man died whose life's work symbolized the consolidation of Islām in everyday life: al-Bukhārī, who produced one of the six collections of *ḥadīth* recognized as authoritative by Jamāʿi-Sunnite Muslims. His fellow collector of *ḥadīth*, Muslim ibn al-Ḥajjāj, died about four years later. About the same time, classical thinkers in other areas of Islāmicate civilization died, among them the great author of *adab*, al-Jāḥiẓ (868/869), the great early ecstatic Ṣūfīs Abu'l Fayḍ Dhu'n-Nūn al-Miṣrī (861) and Abū Yazīd Bisṭāmī (874), the philosopher Yaʿqūb ibn Isḥaq aṣ-Ṣabāḥ al-Kindī (870), and the historian of the conquests al-Balādhurī (*c.* 892). Men of different religious and ethnic heritages, they signified, by the last quarter of the 9th century, the full and varied range of intellectual activities of a civilization that had come of age.

Fragmentation and florescence (870–1041)

THE RISE OF COMPETITIVE REGIONS

The unifying forces operative at the end of the period of conversion and crystallization persisted during the period of fragmentation and florescence; but the caliphal lands in Iraq became less central. Even though Baghdad remained preeminent in cultural prestige, important initiatives were being taken from surrounding "regions": Andalusia; the Maghrib and sub-Saharan Africa; Egypt, Syria, and the holy cities (Mecca and Medina); Iraq; and Iran, Afghanistan, Transoxania, and, toward the end of the period, northern India. Regional courts could compete with the ʿAbbāsids and with each other as patrons of culture. Interregional and intra-regional conflicts were often couched in terms of loyalties formed in the period of conversion and crystallization, but local history provided supplemental identities. Although the ʿAbbāsid caliphate was still a focus of concern and debate, other forms of leadership became important. Just as being Muslim no longer meant being Arab, being cultured no longer meant speaking and writing exclusively in Arabic. Certain Muslims began to cultivate a second language of high culture, New Persian. As in pre-Islāmic times, written as well as spoken bilingualism became important. Ethnic differences were blurred by the effects of peripatetic education and shared languages. Physical mobility was so common that many individuals lived and died far from their places of birth. Cultural creativity was so noticeable that this period is often called the Renaissance of Islām.

New Persian language

Economic changes also promoted regional strengths. Although Baghdad continued to profit from its central location, caliphal neglect of Iraq's irrigation system and southerly shifts in the trans-Asian trade promoted the fortunes of Egypt; the opening of the Sahara to Maghribi Muslims provided a new source of slaves, salt, and minerals; and Egyptian expansion into the Mediterranean opened a major channel for Islāmicate influence on medieval Europe. Islāmdom continued to expand, sometimes as the result of aggression on the part of frontier warriors (*ghāzīs*), but more often as the result of trade. The best symbol of this expansiveness is Ibn Faḍlān, who left a provocative account of his mission in 921, on behalf of the Baghdad caliph, to the Volga Bulgars, among whom he met Swedes coming down the river to trade.

By the beginning of the period of fragmentation and florescence the subject populations of most Muslim rulers were predominantly Muslim, and nonsedentary peoples had ceased to play a major role. The period gave way to a much longer period (dated 1041–1405) in which migratory tribal peoples were once again critically important. In 1041 the reign of the Ghaznavid sultan Masʿūd I ended; by then the Ghaznavid state had lost control over the Seljuq Turks in their eastern Iranian domains and thus inaugurated Islāmdom's second era of tribal expansion. Because localism and cosmopolitanism coexisted in the period of fragmentation and florescence, the period is best approached through a region-by-region survey that underscores phenomena of interregional significance.

ANDALUSIA, THE MAGHRIB, AND SUB-SAHARAN AFRICA

Andalusia, far from the centre of Islāmdom, illustrated the extent of ʿAbbāsid prestige and the assertion of local creativity. In the beginning of the period, Islāmicate rule was represented by the Umayyads at Córdoba; established in 756 by a refugee from the ʿAbbāsid victory over the Syrian Umayyads, the Umayyad dynasty in Córdoba had replaced a string of virtually independent deputies of the Umayyad governors in the Maghrib. At first the Cordoban Umayyads had styled themselves *amīrs*, the title also used by caliphal governors and other local rulers; though refugees from ʿAbbāsid hostility, they continued to mention the ʿAbbāsids in the Friday worship session until 773. Their independence was not made official, however, until their best known member, ʿAbd ar-Raḥmān III (ruled 912–961), adopted the title of caliph in 929 and began having the Friday prayer recited in the name of his own house.

The fact that ʿAbd ar-Raḥmān declared his independence from the ʿAbbāsids while he modeled his court after theirs illustrates the period's cultural complexities. Like the ʿAbbāsids' and the Marwānids', ʿAbd ar-Raḥmān's absolute authority was limited by the nature of his army (Berber tribesmen and Slav slaves) and by his dependence on numerous assistants. His internal problems were compounded by external threats, from the Christian kingdoms in the north and the Fāṭimids in the Maghrib (see below). The Umayyad state continued to be the major Muslim presence in the peninsula until 1010, after which time it became, until 1031, but one of many independent city-states. Nowhere is the connection between fragmentation and florescence more evident than in the courts of these *mulūk al-ṭawāʾif*, or "party kings"; for it was they who patronized some of Andalusia's most brilliant Islāmicate culture. This florescence also demonstrated the permeability of the Muslim–Christian frontier. For example, the poet and theologian Ibn Ḥazm (994–1064) composed love poetry, such as *Ṭawk al-ḥamāmah* (*The Ring of the Dove*), which may have contributed to ideas of chivalric love among the Provençal troubadours.

The flowering of Islāmicate culture in Andalusia

In 870 the Maghrib was divided among several dynasties, all but one of foreign origin, and only one of which, the Aghlabids, nominally represented the ʿAbbāsids. The Muslim Arabs had been very different rulers than any of their predecessors—Phoenicians, Romans, Vandals, or Byzantines—who had occupied but not settled. Their interests in North Africa had been secondary to their objectives in the Mediterranean, so they had restricted themselves to coastal settlements, which they used as staging points for trade with the western Mediterranean or as sources of food for their "metropolitan" population. They had separated themselves from the Berbers with a fortified frontier. The Arabs, however, forced away from the coast in order to compete more effectively with the Byzantines, had quickly tried to incorporate the Berbers, who were also pastoralists. One branch of the Berbers, the Ṣanhājah, extended far into the Sahara, across which they had established a caravan trade with blacks in the Sudanic belt. At some time in the 10th century the Ṣanhājah nominally converted to Islām, and their towns in the Sahara began to assume Muslim characteristics. Around 990 a black kingdom in the Sudan, Ghana, extended itself as far as Audaghost, the Ṣanhājah centre in the Sahara. Thus was black Africa first brought into contact with the Muslim Mediterranean, and thus were the conditions set for dramatic developments in the Maghrib during the 12th and 13th centuries (see below, *Migration and renewal*).

Unification of the Maghrib

In the late 9th century the Maghrib was unified and freed from outside control for the first time. Paradoxically, this independence was achieved by outsiders associated with an international movement of political activism and subversion. Driven underground by ʿAbbāsid intolerance and a maturing ideology of covert revolutionism, the Ismāʿīlī Shīʿites had developed mechanisms to maintain solidarity and undertake political action. These mechanisms can be subsumed under the term *daʿwah*, the same word that had been used for the movement that brought the ʿAbbāsids to power. The *daʿwah*'s ability to communicate rapidly over a large area rested on its traveling operatives as well as on a network of local cells. In the late 9th century an Ismāʿīlī movement, nicknamed the Qarāmiṭah (Qarmatians), had seriously but unsuccessfully threatened the ʿAbbāsids in Syria, Iraq, and Bahrain. Seeking other outlets, a Yemeni operative known as Abū ʿAbd Allāh ash-Shīʿī made contact, on the occasion of the *ḥajj*, with representatives of a Berber tribe that had a history of Kharijite hostility to caliphal control. The *ḥajj* had already become a major vehicle for tying Islāmdom's regions together, and Abū ʿAbd Allāh's movement was only one of many in the Maghrib that would be inaugurated thereby.

In 901 Abū ʿAbd Allāh arrived in the Petite Kabylie (in present-day Algeria); for eight years he prepared for an *imām*, preaching of a millennial restoration of justice after an era of foreign oppression. After conquering the Aghlabid capital al-Qayrawān (in present-day Tunisia), he helped free from a Sijilmassa prison his *imām*, ʿUbayd Allāh, who declared himself the *mahdī*, using a multivalent word that could have quite different meanings for different constituencies. Some Muslims applied *mahdī* to any justice-restoring divinely guided figure; others, including many Jamāʿi-Sunnites, to the apocalyptic figure expected to usher in the millennium before the Last Judgment; and still others, including most Shīʿites, to a returned or restored *imām*. Abū ʿAbd Allāh's followers may have differed in their expectations, but the *mahdī* himself was unequivocal: he was a descendant of ʿAlī and Fāṭimah through Ismāʿīl's disappeared son and therefore was a continuation of the line of the true *imām*. He symbolized his victory by founding a new capital named, after himself, al-Mahdīyah (in present-day Tunisia). During the next half century the "Fāṭimids" tried with limited success to expand westward into the Maghrib and north into the Mediterranean, where they made Sicily a naval base (912–913); but their major goal was Egypt, nominally under ʿAbbāsid control. From Egypt they would challenge the ʿAbbāsid caliphate itself. In 969 the Fāṭimid army conquered the Nile Valley and advanced into Palestine and southern Syria as well.

EGYPT, SYRIA, AND THE HOLY CITIES

The Fāṭimids in Cairo

The Fāṭimids established a new and glorious city, al-Qahirah ("The Victorious"; Cairo), to rival ʿAbbāsid Baghdad. They then adopted the title of caliph, laying claim to be the legitimate rulers of all Muslims as well as head of all Ismāʿīlīs. Now three caliphs reigned in Islāmdom, where there was supposed to be only one. In Cairo the Fāṭimids founded a great mosque–school complex, al-Azhar. They fostered local handicraft production and revitalized the Red Sea route from India to the Mediterranean. They built up a navy to trade as well as to challenge the Byzantines and underscore the ʿAbbāsid caliph's failure to defend and extend the frontiers. Fāṭimid occupation of the holy cities of Mecca and Medina, complete by the end of the 10th century, had economic as well as spiritual significance: it reinforced the caliph's claim to leadership of all Muslims; provided wealth; and helped him keep watch on the West Arabian coast, from the Hejaz to the Yemen, where a sympathetic Zaydī Shīʿite dynasty had ruled since 897. Fāṭimid presence in the Indian Ocean was even strong enough to establish an Ismāʿīlī missionary in Sind. The Fāṭimids patronized the arts; Fāṭimid glass and ceramics were some of Islāmdom's most brilliant. As in other regions, imported styles and tastes were transformed by or supplemented with local artistic impulses, especially in architecture, the most characteristic form of Islāmicate art.

The reign of one of the most unusual Fāṭimid caliphs, al-Ḥākim, from 996 to 1021, again demonstrated the interregional character of the Ismāʿīlī movement. Historians describe al-Ḥākim's personal habits as eccentric, mercurial, and unpredictable to the point of cruelty; his religious values, as inconsistent with official Ismāʿīlī teachings, tending toward some kind of accommodation with the Jamāʿi-Sunnite majority. After he vanished under mysterious circumstances, his religious revisionism was not pursued by his successors or by the Ismāʿīlī establishment in Egypt; but in Syria it inspired a peasant revolt that produced the Druze, who still await al-Ḥākim's return.

When the Fāṭimids expanded into southern Syria, another Shīʿite dynasty, the Ḥamdānid, of Bedouin origin, had been ruling northern Syria from Mosul since 905. In 944 a branch of the family had taken Aleppo; under the leadership of their most famous member, Sayf ad-Dawlah (ruled 945–967), the Ḥamdānids responded aggressively to renewed Byzantine expansionism in eastern Anatolia. They ruled from Aleppo until they were absorbed by the Fāṭimids after 1004; at their court some of Islāmdom's most lastingly illustrious writers found patronage. Two notable examples are the poet al-Mutanabbī (915–965), who illustrated the importance of the poet as a premodern press agent of the court, and al-Fārābī, who tried to reconcile reason and revelation.

Intellectual life under the Ḥamdānids

Al-Fārābī contributed to the ongoing Islāmization of Hellenistic thought. *Falsafah,* the Arabic cognate for the Greek *philosophia,* included metaphysics and logic, as well as the positive sciences, such as mathematics, music, astronomy, and anatomy. *Faylasūfs* often earned their living as physicians, astrologers, or musicians. The *faylasūf*'s whole way of life, like that of the *adīb,* reflected his studies. It was often competitive with that of more self-consciously observant Muslims because the *faylasūf* often questioned the relationship of revelation to real truth. The *faylasūfs* felt free to explore inner truths not exposed to the view of ordinary people; they practiced prudent concealment (*taqīyah*) of their deeper awareness wherever making it public might endanger the social order. The *faylasūfs* shared the principle of concealment with the Shīʿites; both believed, for rather different reasons, that inner truth was accessible to only a very few. This esotericism had counterparts in all premodern societies, where learning and literacy were severely restricted.

IRAQ

Cultural flowering in Iraq. By the late 9th and early 10th centuries the last remnant of the caliphal state was Iraq, under control of the Turkic soldiery. Political decline and instability did not preclude cultural creativity and productivity, however. In fact, Iraq's "generation of 870," loosely construed, contained some of the most striking and lastingly important figures in all of early Islāmicate civilization. Three of them illustrate well the range of culture in late 9th- and early 10th-century Iraq: the historian and Qurʾānic exegete aṭ-Ṭabarī (*c.* 839–923), the theologian Abū al-Ḥasan al-Ashʿarī (*c.* 873–*c.* 935), and the ecstatic mystic al-Ḥallāj (*c.* 858–922).

Abū Jaʿfar Muḥammad ibn Jarīr was born in Ṭabaristān, south of the Caspian Sea, and as a young man he traveled to Baghdad. Rarely could a man earn his living from religious learning; unless he found patronage, he would probably engage in trade or a craft. All the more astounding was the productivity of scholars like aṭ-Ṭabarī, who said that he produced 40 leaves a day for 40 years. The size of his extant works, which include a commentary on the Qurʾān and a universal history, testifies to the accuracy of his claim. His history is unique in sheer size and detail and especially in its long-term impact. His method involved the careful selection, organization, and juxtaposition of separate and often contradictory accounts cast in the form of *ḥadīth.* This technique celebrated the *ummah*'s collective memory and established a range of acceptable disagreement.

At-Ṭabarī

Al-Ashʿarī, from Basra, made his contribution to systematic theological discourse (*kalām*). He had been attracted early to a leading Muʿtazilite teacher, but he broke away at the age of 40. He went on to use Muʿtazilite methods of reasoning to defend popular ideas such as the eternality and literal truth of the Qurʾān, and the centrality of Muḥammad's *sunnah* as conveyed by the *ḥadīth.* Where his approach yielded objectionable results, such as an anthropomorphic rendering of God or a potentially polytheistic understanding of his attributes, al-Ashʿarī resorted to the principle of *bilā kayfah* ("without regard to the how"), whereby a person of faith accepts that certain fundamentals are true without regard to how they are true and that divine intention is not always accessible to human intelligence. Al-Ashʿarī's harmonization also produced a simple creed, which expressed faith in God, his angels, and his books, and affirmed belief in Muḥammad as God's last messenger and in the reality of death, physical resurrection, the Last Judgment, and heaven and hell. Taken together, aṭ-Ṭabarī's historiography and al-Ashʿarī's theology symbolize the consolidation of Jamāʿi-Sunnite, Sharīʿah-minded thought and piety.

Al-Ashʿarī

The most visible and powerful 10th-century exponent of Ṣūfism was al-Ḥallāj. By his day, Ṣūfism had grown far beyond its early forms, which were represented by al-Ḥasan al-Baṣrī (died 728), who practiced *zuhd,* or rejection of the world, and by Rābiʿah al-ʿAdawīyah (died 801), who formulated the Ṣūfī ideal of a disinterested love of God. The mystics Abū Yazīd Bisṭāmī (died 874) and al-Junayd (died 910) had begun to pursue the experience of unity with God, first by being "drunk" with his love and with love of him, and then by acquiring life-transforming self-possession and control. Masters (called *shaykh*s or *pīr*s) were beginning to attract disciples (*murīd*s) to their way. Like other Muslims who tried to go "beyond" the Sharīʿah to inner truth, the Ṣūfīs practiced concealment of inner awareness (*taqīyah*). Al-Ḥallāj, one of al-Junayd's disciples, began to travel and preach publicly, however. His success was disturbing enough for the authorities in Baghdad to have him arrested and condemned to death; he was tortured and beheaded, and finally his body was burned. Yet his career had shown the power of Ṣūfism, which would by the 12th century become an institutionalized form of Islāmic piety.

Al-Ḥallāj

The Būyid dynasty. Long before, however, a major political change occurred at Baghdad. In 945 control over the caliphs passed from their Turkish soldiery to a dynasty known as the Būyids or Buwayhids. The Būyids came from Daylam, near the southern coast of the Caspian Sea. Living beyond the reach of the caliphs in Baghdad, its residents had identified with Imāmī Shīʿism. By about 930, three sons of a fisherman named Būyeh had emerged as leaders in Daylam. One of them conquered Baghdad, not replacing the caliph but ruling in his name. The fact that they were Shīʿite, as were the Idrīsids, Fāṭimids, and Ḥamdānids, led scholars to refer to the period from the mid-10th to mid-11th century as the Shīʿite century.

Like other contemporary rulers, the Būyids were patrons of culture, especially of speculative thought (Shīʿism, Muʿtazilism, *kalām,* and *falsafah*). Jamāʿi-Sunnite learning continued to be patronized by the caliphs and their families. The Būyids favoured no one party over another. However, their openness paradoxically invited a hardening in Jamāʿi-Sunnite thought. Būyid attempts to maintain the cultural brilliance of the court at Baghdad were limited by a decline in revenue occasioned partly by a shift in trade routes to Fāṭimid Egypt, and partly by long-term neglect of Iraq's irrigation works. The caliphs had occasionally made land assignments (*iqṭāʿ*s) to soldiers in lieu of paying salaries; now the Būyids extended the practice to other individuals and thus removed an important source of revenue from central control. After 983, Būyid territo-

Būyid patronage of culture

ries were split among various members of the family, and pressure was applied to their borders from both the west (by Ḥamdānids and Fāṭimids) and the east (by Sāmānids, Ghaznavids, and Seljuqs; see below).

The economic difficulties of Būyid Iraq promoted urban unrest, accounts of which provide a rare glimpse into the lives of ordinary Muslim town dwellers. Numerous movements served as outlets for socioeconomic grievances, directed most often toward the wealthy or the military. The concentration of wealth in the cities had produced a bipolar stratification system conveyed in the sources by a pair of words, *khāṣṣ* (special) and *ʿāmm* (ordinary). In the environment of 10th- and 11th-century Iraq, an instance of rising food prices or official maltreatment could easily spark riots of varying size, duration, and intensity. Strategies for protest included raiding, looting, and assault. Some movements were more coherently ideological than others, and various forms of piety could reflect socioeconomic distinctions. Some movements were particularly attractive to artisans, servants, and soldiers, as was the case with the proponents of Ḥadīth, whose mentor, Aḥmad ibn Ḥanbal (died 855), was viewed as a martyr because of his suffering at the hands of the Caliph. Other forms of piety, such as Shīʿism, could be associated with wealthier elements among the landowning and merchant classes.

Social action associations: *ʿayyār* and *futūwah*

Beneath the more organized forms of social action lay a more fluid kind of association, most often described by the labels *ʿayyār* and *futūwah.* These terms refer to individuals acting in concert, as needed, on the basis of certain rough-hewn concepts of proper male public behaviour. Such associations had counterparts in the late Hellenistic world, just as they have parallels in the voluntary protective associations formed in the 19th and 20th centuries whenever official institutions of protection have been either chronically or temporarily deficient. For some of the Islāmicate "gangs" or "clubs," thuggery may have been the norm; for others, the figure of the fourth caliph and first *imām,* ʿAlī, seems to have provided an exemplar. Even though Shīʿites had become a separate group with a distinctive interpretation of ʿAlī's significance, a more generalized affection for the family of the Prophet, and especially for ʿAlī, was widespread among Jamāʿi-Sunnites. ʿAlī had come to be recognized as the archetypal young male (*fatā*); a related word, *futūwah,* signified groups of young men who pursued such virtures as courage, aiding the weak, generosity, endurance of suffering, love of truth, and hospitality.

Premodern Islāmicate societies were characterized by a high degree of fluidity, occasionalism, and voluntarism in the structuring of associations, organizations, loyalties, and occupations. Although all societies must develop ways to maintain social boundaries, ease interaction among groups, and buffer friction, the ways in which Muslim societies have fulfilled these needs seem unusually difficult to delineate. For example, in Muslim cities of the period under discussion, the only official officeholders were appointees of the central government, such as the governor; the *muḥtasib,* a transformed Byzantine *agoranomos* who was monitor of public morality as well as of fair-market practice; or the *sahib ash-shurtah,* head of the police. In the absence of an organized church or ordained clergy, those whose influence derived from piety or learning were influential because they were recognized as such, not because they were appointed; and men of very different degrees of learning might earn the designation of *ʿālim.* Although the ruler was expected to contribute to the maintenance of public services, neither he nor anyone else was obligated to do so. Though the ruler might maintain prisons for those whose behaviour he disapproved, the local *qāḍī*s had need of none, relying generally on persuasion or negotiation and borrowing the caliphal police on the relatively rare occasion on which someone needed to be brought before them by force. There was no formalized mode of succession for any of the dynasties of the time. Competition, sometimes armed, was relied upon to produce the most qualified candidate.

Patronage was an important basis of social organization. The family served as a premodern welfare agency; where it was absent, minimal public institutions, such as hospitals, provided. One of the most important funding mechanisms for public services was a private one, the *waqf.* The *waqf* provided a legal way to circumvent the Sharīʿah's requirement that an individual's estate be divided among many heirs. Through a *waqf,* an individual could endow an institution or group with all or part of his estate, in perpetuity, before his death. A *waqf* might provide books for a school, candles or mats for a mosque, salaries of religious functionaries, or land for a hospital or caravansary. *Waqf* money or lands were indivisible, although they might contribute to the welfare of a potential heir who happened to be involved in the *waqf*-supported activity. The *waqf,* like other forms of patronage, provided needed social services without official intervention. On other occasions, wealthy individuals, especially those connected with the ruling family, might simply patronize favourite activities. In addition to patronage, many other overlapping ties bound individual Muslims together: loyalties to an occupation—soldier, merchant, learned man, artisan, government worker; loyalties to a town or neighbourhood, or to a form of piety, or to persons to whom one made an oath for a specific purpose; and ties to patron or to family, especially foster-parentage (*istināʿ*), the counterpart of which was significant in medieval Christendom.

The relationship of individual and group action

The Qurʾān and Sharīʿah discouraged corporate responsibility in favour of individual action; even the legal scope of partnership was limited. Yet the unstable political realities that had militated against the emergence of broad-based institutions sometimes called for corporate action, as when a city came to terms with a new ruler or invader. In those cases, a vaguely defined group of notables, known usually as *aʿyān,* might come together to represent their city in negotiations, only to cease corporate action when the more functional small-group loyalties could safely be resumed. Within this shifting frame of individuals and groups, the ruler was expected to maintain a workable, if not equitable, balance. More often than not the real ruler was a local *amīr* of some sort. For this reason, the de facto system of rule that emerged during this period, despite the persistence of the central caliphate in Baghdad, has sometimes been referred to as the *aʿyān–amīr* system.

The city's physical and social organization reflected this complex relationship between public and private, and between individual and group: physically separated quarters; multiple markets and mosques; mazelike patterns of narrow streets and alleys with dwellings oriented toward an inner courtyard; an absence of public meeting places other than bath, market, and mosque; and the concentration of social life in private residences. The *qāḍī* and *adīb* at-Tanūkhī provides a lively and humorous picture of 10th-century Baghdad, of a society of individuals with overlapping affiliations and shifting statuses: saints and scoundrels, heroes and rogues, rich men and poor. This mobility is illustrated by at-Tanūkhī's boast to a rival, "My line begins with me while yours ends with you." The prose genre of *maqāmah,* said to have been invented by al-Hamadhānī (died 1007), recounted the exploits of a clever, articulate scoundrel dependent on his own wits for his survival and success.

IRAN, AFGHANISTAN, AND INDIA

In the middle of the "Shīʿite century" a major Sunnite revival occurred in eastern Islāmdom in connection with the emergence of the second major language of Islāmicate high culture, New Persian. This double revival was accomplished by two Iranian dynasties, the Sāmānids and the Ghaznavids; Ghaznavid zeal even spilled over into India.

The Sāmānids. The Sāmānid dynasty (819–999) stemmed from a local family appointed by the ʿAbbāsids to govern at Bukhara and Samarkand. Gradually the Sāmānids had absorbed the domains of the rebellious Ṭāhirids and Ṣaffārids in northeastern Iran and reduced the Ṣaffārids to a small state in Sīstān. The Sāmānids, relying on Turkic slave troops, also managed to contain the migratory pastoralist Turkic tribes who continually pressed on Iran from across the Oxus River. In the 950s they even managed to convert some of these Turkic tribes to Islām.

The Sāmānid court at Bukhara attracted leading scholars, such as the philosophers Abū Bakr ar-Rāzī (died 925) and Avicenna (Ibn Sīnā; 980–1037), who later worked for the Būyids; and the poet Ferdowsī (died *c.* 1020). Though not Shī'ites, the Sāmānids expressed an interest in Shī'ite thought, especially in its Ismā'īlī form, which was then the locus of so much intellectual vitality. The Sāmānids also fostered the development of a second Islāmicate language of high culture, New Persian. It combined the grammatical structure and vocabulary of spoken Persian with vocabulary from Arabic, the existing language of high culture in Iran. A landmark of this "Persianizing" of Iran was Ferdowsī's epic poem, the *Shāh-nāmeh* ("Book of Kings"), written entirely in New Persian in a long-couplet form (*masnavī*) derived from Arabic. Covering several thousand years of detailed mythic Iranian history, Ferdowsī brought Iran's ancient heroic lore, and its hero Rustam, into Islāmicate literature and into the identity of self-consciously Iranian Muslims. He began to compose the poem under the rule of the Sāmānids; but he dedicated the finished work to a dynasty that had meanwhile replaced them, the Ghaznavids.

The Ghaznavids. The Ghaznavid dynasty was born in a way that had become routine for Islāmicate polities. Sebüktigin (ruled 977–997), a Sāmānid Turkic slave governor in Ghazna (now Ghaznī), in the Afghan mountains, made himself independent of his masters as their central power declined. His eldest son, Maḥmūd, expanded into Būyid territory in western Iran, identifying himself staunchly with Sunnite Islām. Presenting himself as a frontier warrior against the pagans, Maḥmūd invaded and plundered northwestern India, establishing a permanent rule in the Punjab; but it was through ruling Iran, which gave a Muslim ruler true prestige, that Maḥmūd sought to establish himself. He declared his loyalty to the 'Abbāsid caliph, whose "investiture" he sought, and expressed his intention to defend Sunnite Islām against the Shī'ite Būyids. Although he and his regime were proud of their Turkic descent, Maḥmūd encouraged the use of New Persian, with its echoes of pre-Islāmic Iranian glory, for administration and for prose as well as poetry. This combination of Turkic identity and Persian language would characterize and empower many other Muslim rulers.

Sebüktigin

To Ghazna Maḥmūd brought, sometimes by force, writers and artisans who could adorn his court. Among these was al-Bīrūnī (973–*c.* 1050), whose scholarly achievements no contemporary could rival. Before being brought to Ghazna, al-Bīrūnī had served the Sāmānids and the Khwārazm-Shāhs, a local dynasty situated just west of the Oxus River. Al-Bīrūnī's works included studies of astronomy (he even suggested a heliocentric universe), gems, drugs, mathematics, and physics; but his most famous book, inspired by accompanying Maḥmūd on his Indian campaigns, was a survey of Indian life, language, religion, and culture.

Like most other rulers of the day, Maḥmūd styled himself *amīr* and emphasized his loyalty to the caliph in Baghdad; but he and later Ghaznavid rulers also called themselves by the Arabic word, *sulṭān* (sultan). Over the next five centuries the office of sultan would become an alternative to caliph. The Ghaznavid state presaged other changes as well, especially by stressing the cleavage between ruler and ruled and by drawing into the ruling class not only the military but also the bureaucracy and the learned establishment. So tied was the ruling establishment to the ruler that it even moved with him on campaign. Ghaznavid "political theory" shared with other states the concept of the circle of justice or circle of power; *i.e.,* that justice is best preserved by an absolute monarch completely outside society; that such a ruler needs an absolutely loyal army; and that maintaining such an army requires prosperity, which in turn depends on the good management of an absolute ruler.

Bu'l-Fazl-i Bayhaqī (995–1077) worked in the Ghaznavid chancery and wrote a remarkable history of the Ghaznavids, the first major prose work in New Persian. He exhibited the broad learning of even a relatively minor figure at court; in his history he combined the effective writing skills of the chancery employee, the special knowledge of Qur'ān and *ḥadīth,* and the sophisticated and entertaining literature—history, poetry, and folklore—that characterized the *adīb.* He provided a vivid picture of life at court, graphically portraying the pitfalls of military absolutism—the dependence of the monarch on a fractious military and a large circle of assistants and advisors, who could mislead him and affect his decision making through internecine maneuvering and competition. In the reign of Maḥmūd's son, Mas'ūd I, the weaknesses in the system had already become glaringly apparent. At the Battle of Dandānqān (1040), Mas'ūd lost control of Khorāsān, his main holding in Iran, to the pastoralist Seljuq Turks; he then decided to withdraw to Lahore in his Indian domains, from which his successors ruled until overtaken by the Ghūrids in 1186.

THE DECLINE OF THE CALIPHATE AND RISE OF EMIRATES

By the end of Mas'ūd's reign, government in Islāmdom had become government by *amir.* Caliphal centralization had lasted 200 years; and even after the caliphal empire became too large and complex to be ruled from a single centre, the separate emirates that replaced it all defined their legitimacy in relation to it, for or against. In fact, the caliphate's first systematic description and justification was undertaken just when its impracticality was being demonstrated. As the Ghaznavids were ruling in Iran as "appointed" defenders of the caliph, a Baghdadi legal scholar named al-Māwardī retrospectively delineated the minimal requirements of the caliphate and tried to explain why it had become necessary for caliphal powers to be "delegated" in order for the *ummah*'s security to be maintained. Whereas earlier legists had tied the caliph's legitimacy to his defense of the borders, al-Māwardī separated the two, maintaining the caliph as the ultimate source of legitimacy and the guardian of pan-Islāmic concerns, and relegating day-to-day government to his "appointees." Al-Māwardī may have hoped that the Ghaznavids would expand far enough to be "invited" by the caliph to replace the uninvited Shī'ite Būyids. This replacement did occur, three years before al-Māwardī's death; however, it was not the Ghaznavids who appeared in Baghdad but rather the migratory pastoralist Turks who had meanwhile replaced them. The Seljuqs joined many other migrating groups to produce the next phase of Islāmicate history.

Migration and renewal (1041–1405)

During this period, migrating peoples once again played a major role, perhaps greater than that of the Arabs during the 7th and 8th centuries. No other civilization in premodern history experienced so much in-migration, especially of alien and disruptive peoples, or showed a greater ability to assimilate as well as to learn from outsiders. Nowhere has the capacity of a culture to redefine and incorporate the strange and the foreign been more evident. In this period, which ends with the death in 1405 of Timur (Tamerlane), the last great tribal conqueror, the tense yet creative relationship between sedentary and migratory peoples emerged as one of the great themes of Islāmicate history, played out as it was in the centre of the great arid zone of Eurasia. Because this period can be seen as the history of peoples as well as of regions, and because the mobility of those peoples brought them to more than one cultural region, this period should be treated group by group rather than region by region.

Relationship of sedentary and migratory peoples

As a general term "migrating" peoples is preferable because it does not imply aimlessness, as "nomadic" does; or herding, as "pastoralist" does; or kin-related, as "tribal" does. "Migrating" focuses simply on movement from one home to another. Although the Franks, as the crusaders are called in Muslim sources, differed from other migrating peoples, most of whom were pastoralists related by kinship, they too were migrating warriors organized to invade and occupy peoples to whom they were hostile and alien. Though not literally tribal, they appeared to behave like a tribe with a distinctive way of life and a solidarity based on common values, language, and objectives. Viewing them as alien immigrants comparable to, say, the Mongols, helps to explain their reception: how they came to be assimilated into the local culture and drawn into the intra-Muslim factional competition and fighting that was under way in Syria when they arrived.

TURKS

For almost 400 years a succession of Turkic peoples entered eastern Islāmdom from Central Asia. These nearly continuous migrations can be divided into three phases: Seljuqs (1055–92), Mongols (1256–1411), and neo-Mongols (1369–1405). Their long-term impact, more constructive than destructive on balance, can still be felt through the lingering heritage of the great Muslim empires they inspired. The addition of tribally organized warrior Turks to the already widely used Turkic slave soldiery gave a single ethnic group an extensive role in widening the gap between rulers and ruled.

Seljuq Turks. The Seljuqs were a family among the Oğuz Turks, a label applied to the migratory pastoralists of the Syrdarya–Oxus basin. Their name has come to stand for the group of Oğuz families led into Ghaznavid Khorāsān after they had been converted to Sunnite Islām, probably by Ṣūfī missionaries after the beginning of the 11th century. In 1040 the Seljuqs' defeat of the Ghaznavid sultan allowed them to proclaim themselves rulers of Khorāsān. Having expanded into western Iran as well, Toghrïl Beg, also using the title "sultan," was able to occupy Baghdad (1055) after "petitioning" the ʿAbbāsid caliph for permission. The Seljuqs quickly took the remaining Būyid territory and began to occupy Syria, whereupon they encountered Byzantine resistance in the Armenian highlands. In 1071 a Seljuq army under Alp-Arslan defeated the Byzantines at Manzikert north of Lake Van; while the main Seljuq army replaced the Fāṭimids in Syria, large independent tribal bands occupied Anatolia, coming closer to the Byzantine capital than had any other Muslim force.

Policies of Niẓām al-Mulk. The Seljuqs derived their legitimacy from investiture by the caliph, and from "helping" him reunite the *ummah;* yet their governing style prefigured the emergence of true alternatives to the caliphate. Some of their Iranian advisers urged them to restore centralized absolutism as it had existed in pre-Islāmic times and in the period of Marwānid–ʿAbbāsid strength. The best known proponent was Niẓām al-Mulk, chief minister to the second and third Seljuq sultans, Alp-Arslan and Malik-Shāh. Niẓām al-Mulk explained his plans in his *Seyāsat-nāmeh,* one of the best known manuals of Islāmicate political theory and administration. He was unable, however, to persuade the Seljuq sultans to assert enough power over other tribal leaders. Eventually the Seljuq sultans, like so many rulers before them, alienated their tribal supporters and resorted to the costly alternative of a Turkic slave core, whose leading members were appointed to tutor and train young princes of the Seljuq family to compete for rule on the death of the reigning sultan. The tutors were known as *atabeg*s; more often than not, they became the actual rulers of the domains assigned to their young charges, cooperating with urban notables (*aʿyān*) in day-to-day administration.

Founding of *madrasah*s

Although Niẓām al-Mulk was not immediately successful, he did contribute to long-term change. He encouraged the establishment of state-supported schools (*madrasah*s); those he personally patronized were called Niẓāmīyahs. The most important Niẓāmīyah was founded in Baghdad in 1067; here Niẓām al-Mulk gave government stipends to teachers and students whom he hoped he could subsequently not only appoint to the position of *qāḍī* but also recruit for the bureaucracy. Systematic and broad instruction in Jamāʿi-Sunnite learning would counteract the disruptive influences of non-Sunnite or anti-Sunnite thought and activity, particularly the continuing agitation of Ismāʿīlī Muslims. In 1090 a group of Ismāʿīlīs established themselves in a mountain fortress at Alamūt in the mountains of Daylam. From there they began to coordinate revolts all over Seljuq domains. Nominally loyal to the Fāṭimid caliph in Cairo, the eastern Ismāʿīlīs confirmed their growing independence and radicalism by supporting a failed contender for the Fāṭimid caliphate, Nizār. For that act they were known as the Nizārī Ismāʿīlīs. They were led by Ḥasan-e Ṣabbāḥ and were dubbed by their detractors the *hashīshīyah* (assassins) because they practiced political murder while they were allegedly under the influence of hashish.

Niẓām al-Mulk's *madrasah* system enhanced the prestige and solidarity of the Jamāʿi-Sunnite *ʿulamāʾ* without actually drawing them into the bureaucracy or combating anti-Sunnite agitation, but it also undermined their autonomy. It established the connection between state-supported education and office holding, and it subordinated the spiritual power and prestige of the *ʿulamāʾ* to the indispensable physical force of the military *amīr*s. Niẓām al-Mulk unintentionally encouraged the independence of these *amīr*s by extending the *iqṭāʿ* system beyond Būyid practice; he regularly assigned land revenues to individual military officers, assuming that he could keep them under bureaucratic control. When that failed, his system increased the *amīr*s' independence and drained the central treasury.

Al-Ghazālī's teaching

The *madrasah* system had other unpredictable results that can be illustrated by al-Ghazālī, who was born in 1058 at Ṭūs and in 1091 was made head of the Baghdad Niẓāmīyah. For four years, to great admiration, he taught both *fiqh* and *kalam* and delivered critiques of *falsafah* and Ismāʿīlī thought. According to his autobiographical work *Al-Munqidh min aḍ-ḍalāl* (*The Deliverer from Error*), the more he taught, the more he doubted, until his will and voice became paralyzed. In 1095 he retreated from public life, attempting to arrive at a more satisfying faith. He undertook a radically skeptical reexamination of all of the paths available to the pious Muslim, culminating in an incorporation of the active, immediate, and inspired experience of the Ṣūfīs into the Sharīʿah-ordered piety of the public cult. For his accomplishments, al-Ghazālī was viewed as a renewer (*mujaddid*), a role expected by many Muslims to be filled by at least one figure at the turn of every Muslim century.

Ṭarīqah fellowships. In the 12th century Muslims began to group themselves into *ṭarīqah*s, fellowships organized around and named for the *ṭarīqah* ("way" or "path") of given masters. Al-Ghazālī may have had such a following himself. One of the first large-scale orders, the Qādirīyah, formed around the teachings of ʿAbd al-Qādir al-Jīlānī of Baghdad. Though rarely monastic in the European sense, the activities of a *ṭarīqah* often centred around assembly halls (called *khānqāh, zāwīyah,* or *tekke*) that could serve as places of retreat or accommodate special spiritual exercises. The *dhikr,* for example, is a ceremony in which devotees meditated on the name of God to the accompaniment of breathing exercises, music, or movement, so as to attain a state of consciousness productive of a sense of union with God. Although shortcuts and excesses have often made Ṣūfism vulnerable to criticism, its most serious practitioners have conceived of it as a disciplined extension of Sharīʿah-minded piety, not an escape. In fact, many Ṣūfīs have begun their path through supererogatory fulfillment of standard ritual requirements.

Thousands of *ṭarīqah*s sprang up over the centuries, some associated with particular occupations, locales, or classes. It is possible that by the 18th century most adult Muslim males had some connection with one or more *ṭarīqah*s. The structure of the *ṭarīqah* ensued from the charismatic authority of the master, who, though not a prophet, replicated the direct intimacy that the prophets had shared with God. This quality he passed on to his disciples through a hierarchically ordered network that could extend over thousands of miles. The *ṭarīqah*s thus became powerful centripetal forces among societies in which formal organizations were rare; but the role of the master became controversial because followers often made saints or intercessors of especially powerful Ṣūfī leaders and made shrines or pilgrimage sites of their tombs or birthplaces. Long before these developments could combine to produce stable alternatives to the caliphal system, Seljuq power had begun to decline, only to be replaced for a century and a half with a plethora of small military states. When the Frankish crusaders arrived in the Holy Land in 1099, no one could prevent them from quickly establishing themselves along the eastern Mediterranean coast.

FRANKS

The call for the Crusades. At the Council of Clermont in 1095 Pope Urban II responded to an appeal from the

Byzantine emperor for help against the Seljuq Turks, who had expanded into western Anatolia just as the Kipchak Turks in the Ukraine had cut off newly Christian Russia from Byzantium. The First Crusade, begun the next year, brought about the conquest of Jerusalem in 1099. The Christian Reconquista (reconquest) of Spain was already under way, having scored its first great victory at Toledo in 1085. Ironically, modern historiography has concentrated on the crusades that failed and virtually ignored the ones that succeeded. In the four centuries between the fall of Toledo and the fall of Granada (1492), Spanish Christians replaced Muslim rulers throughout the Iberian Peninsula, although Muslims remained as a minority under Christian rule until the early 17th century. In the 200 years from the fall of Jerusalem to the end of the Eighth Crusade (1291), western European crusaders failed to halt the Turkish advance or to establish a permanent presence in the Holy Land. By 1187 local Muslims had managed to retake Jerusalem and thereby contain Christian ambitions permanently. By the time of the Fourth Crusade (1202–04) the crusading movement had been turned inward against Christian heretics such as the Byzantines.

Successful crusades in the Iberian Peninsula

Effect of the Crusades in Syria. The direct impact of the Crusades on Islāmdom was limited largely to Syria. For the century during which western European Christians were a serious presence there, they were confined to their massive coastal fortifications. The crusaders had arrived in Syria at one of its most factionalized periods prior to the 20th century. Seljuq control, never strong, was then insignificant; local Muslim rule was anarchic; the Seljuq regime in Baghdad was competing with the Fāṭimid regime in Egypt; and all parties in Syria were the target of the Nizārī Ismāʿīlī movement at Alamūt. The crusaders soon found it difficult to operate as more than just another faction. Yet the significance of the crusaders as a force against which to be rallied should not be underestimated any more than should the significance of Islāmdom as a force against which Christendom could unite.

The crusaders' situation encouraged interaction with the local population and even assimilation. They needed the food, supplies, and services available in the Muslim towns. Like their Christian counterparts in Spain, they took advantage of the enemy's superior skills, in medicine and hygiene, for example. Because warfare was seasonal and occasional, they spent much of their time in peaceful interaction with their non-Christian counterparts. Some early-generation crusaders intermarried with Arab Muslims or Arab Christians and adopted their personal habits and tastes, much to the dismay of Christian latecomers. An intriguing account of life in Syria during the Crusades can be found in the *Kitāb al-Iʿtibār* ("Book of Reflection"), the memoirs of Usamah ibn Munqidh (1095–1188). Born in Syria, he was a small boy when the first generation of Franks controlled Jerusalem. As an adult he fought with Saladin (see below) and lived to see him unite Egypt with Syria and restore Jerusalem to Muslim control. In this fine example of Islāmicate autobiographical writing, Usamah draws a picture of the Crusades not easily found in European sources: Christians and Muslims observing, and sometimes admiring, each others' skills and habits, from the battlefield to the bathhouse. Although the Franks in Syria were clearly influenced by the Muslims, the Crusades seem to have contributed relatively little to the overall impact of Islāmicate culture on Europe, even though they constituted the most prolonged direct contact.

Although the crusaders never formed a united front against the Muslims, Syrian Muslims did eventually form a united front against them, largely through the efforts of the family of the *amīr* Zangī, a Turkic slave officer appointed Seljuq representative in Mosul in 1127. After Zangī had extended his control through northern Syria, one of his sons and successors, Nureddin (Nūr ad-Dīn), based at Aleppo, was able to tie Zangī's movement to the frontier warrior (*ghāzī*) spirit. This he used to draw together urban and military support for a *jihād* against the Christians. After taking Damascus, he established a second base in Egypt. He offered help to the failing Fāṭimid regime in return for being allowed to place one of his own lieutenants, Saladin (Ṣalāḥ ad-Dīn Yūsuf ibn Ayyūb), as chief minister to the Fāṭimid caliph, thus warding off a crusader alliance with the Fāṭimids. This action gave Nureddin two fronts from which to counteract the superior seaborne and naval support the crusaders were receiving from western Europe and the Italian city-states. Three years before Nureddin's death in 1174, Saladin substituted himself for the Fāṭimid caliph he theoretically served, thus ending more than 200 years of Fāṭimid rule in Egypt. When Nureddin died, Saladin succeeded him as head of the whole movement. When Saladin died in 1193, he had recaptured Jerusalem (1187) and begun the reunification of Egypt and Syria; his successors were known, after his patronymic, as the Ayyūbids. The efforts of a contemporary ʿAbbāsid caliph, an-Nāṣir, to revive the caliphate seem pale by comparison.

The Ayyūbids ruled in Egypt and Syria until 1250, when they were replaced first in Egypt and later in Syria by the leaders of their own slave-soldier corps, the Mamlūks. It was they who expelled the remaining crusaders from Syria, subdued the remaining Nizārī Ismāʿīlīs there, and consolidated Ayyūbid holdings into a centralized state. That state became strong enough in its first decade to do what no other Muslim power could: in 1260 at ʿAyn Jālūt, south of Damascus, the Mamlūk army defeated the recently arrived Mongols and expelled them from Syria.

Final expulsion of crusaders by Ayyūbids

MONGOLS

The Mongols were pagan, horse-riding tribes of the north-eastern steppes of Central Asia. In the early 13th century, under the leadership of Genghis Khan, they formed, led, and gave their name to a confederation of Turkic tribes that they channeled into a movement of global expansion, spreading east into China, north into Russia, and west into Islāmdom. Like other migratory peoples before them, Arabs, Berbers, and Turks, they had come to be involved in citied life through their role in the caravan trade. Unlike others, however, they did not convert to Islām before their arrival. Furthermore, they brought a greater hostility to sedentary civilization, a more ferocious military force, a more cumbersome material culture, a more complicated and hierarchical social structure, and a more coherent sense of tribal law. Their initial impact was physically more destructive than that of previous invaders, and their long-term impact perhaps more socially and politically creative.

First Mongol incursions. The first Mongol incursions into Islāmdom in 1220 were a response to a challenge from the Khwārezm-Shāh ʿAlāʾ ad-Dīn Muḥammad, the aggressive reigning leader of a dynasty formed in the Oxus Delta by a local governor who had rebelled against the Seljuq regime in Khorāsān. Under Genghis Khan's leadership, Mongol forces destroyed numerous cities in Transoxania and Khorāsān in an unprecedented display of terror and annihilation. By the time of Genghis Khan's death in 1227, his empire stretched from the Caspian Sea to the Sea of Japan. A later successor, Möngke, decided to extend the empire in two new directions. From the Mongol capital of Karakorum, he simultaneously dispatched Kublai Khan to southern China (where Islām subsequently began to expand inland) and Hülegü to Iran (1256). Hülegü had already received Sunnite ambassadors who encouraged him to destroy the Ismāʿīlī state at Alamūt; this he did and more, reaching Baghdad in 1258, where he terminated and replaced the caliphate. The ʿAbbāsid line continued, however, until 1517; the Mamlūk sultan Baybars I, shortly after his defeat of the Mongols, invited a member of the ʿAbbāsid house to "invest" him and to live in Cairo as spiritual head of all Muslims.

The Mongol regimes in Islāmdom quickly became rivals. The Il-Khans controlled the Tigris–Euphrates valley and Iran; the Chagatai dominated the Syrdarya and Oxus basins, the Kābul mountains, and eventually the Punjab; and the Golden Horde was concentrated in the Volga basin. The Il-Khans ruled in the territories where Islām was most firmly established. They patronized learning of all types and scholars from all parts of the vast Mongol empire, especially China. Evincing a special interest in nature, they built a major observatory in Azerbaijan. Just as enthusiastically as they had destroyed citied life, they now rebuilt it, relying as had all previous invaders of Iran on the administrative skills of indigenous Persian-speaking bureaucrats. The writings of one of these men, ʿAṭā Malek Joveynī, who was appointed Mongol governor in Baghdad in 1259, described the type of rule the Mongols sought to impose. It has been called the military patronage state because it involved a reciprocal relationship between the foreign tribal military conquerors and their subjects. The entire state was defined as a single mobile military force connected to the household of the monarch; with no fixed capital, it moved with the monarch. All non-Turkic state workers, bureaucratic or religious, even though not military specialists, were defined as part of the army (*asker*); the rest of the subject population, as the herds (*raʿīyah*). The leading tribal families could dispose of the wealth of the conquered populations as they wished, except that their natural superiority obligated them to reciprocate by patronizing whatever of excellence the cities could produce. What the Ghaznavids and Seljuqs had begun, the Mongols now accomplished. The self-confidence and superiority of the leading families were bolstered by a fairly elaborate set of tribal laws, inherited from Genghis Khan and known as the Yasa, which served to regulate personal status and criminal liability among the Mongol elite, as did the Sharīʿah among Muslims. In Il-Khanid hands, this dynastic law merely coexisted but did not compete with Sharīʿah; but in later Turkic regimes a reconciliation was achieved that extended the power of the rulers beyond the limitations of an autonomous Sharīʿah.

Military patronage system of Mongol rule

Conversion of Mongols to Islām. For a time the Il-Khans tolerated and patronized all religious persuasions, Sunnite, Shīʿite, Buddhist, Nestorian Christian, Jewish, and pagan. But in 1295 a Buddhist named Maḥmūd Ghāzān became Khan and declared himself Muslim, compelling other Mongol notables to follow suit. His patronage of Islāmicate learning fostered such brilliant writers as Rashīd ad-Dīn, the physician and scholar who authored one of the most famous Persian universal histories of all time. The Mongols, like other Islāmicate dynasties swept into power by a tribal confederation, were able to unify their domains for only a few generations. By the 1330s their rule had begun to be fragmented among myriad local leaders. Meanwhile, on both Mongol flanks, other Turkic Muslim powers were increasing in strength.

To the east the Delhi Sultanate of Turkic slave soldiers withstood Mongol pressure, benefited from the presence of scholars and administrators fleeing Mongol destruction, and gradually began to extend Muslim control south into India, a feat that was virtually accomplished under Muḥammad ibn Tughluq. Muslim Delhi was a culturally lively place that attracted a variety of unusual persons. Muḥammad ibn Tughluq himself was, like many later Indian Muslim rulers, well read in philosophy, science, and religion. Not possessing the kind of dynastic legitimacy the pastoralist Mongols had asserted, he tied his legitimacy to his support for the Sharīʿah, and he even sought to have himself invested by the ʿAbbāsid "caliph" whom the Mamlūks had taken to Cairo. His concern with the Sharīʿah coincided with the growing popularity of Ṣūfism, especially as represented by the massive Chishti *ṭarīqah*. Its most famous leader, Niẓām ad-Dīn Awliyāʾ, had been a spiritual adviser to many figures at court before Muḥammad ibn Tughluq came to the throne, as well as to individual Hindus and Muslims alike. In India, Ṣūfism, which inherently undermined communalism, was bringing members of different religious communities together in ways very rare in the more westerly parts of Islāmdom.

To the west, the similarly constituted Mamlūk state continued to resist Mongol expansion. Its sultans were chosen, on a nonhereditary basis, from among a group of freed slaves who acted as the leaders of the various slave corps. At the death of one sultan the various military corps would compete to see whose leader would become the next sultan. The leaders of the various slave corps formed an oligarchy that exercised control over the sultan. Although political instability was the frequent and natural result of such a system, cultural florescence did occur. The sultans actively encouraged trade and building, and Mamlūk Cairo became a place of splendour, filled with numerous architectural monuments. While the Persian language was becoming the language of administration and high culture over much of Islāmdom, Arabic alone continued to be cultivated in Mamlūk domains, to the benefit of a diversified intellectual life. Ibn an-Nafīs (died 1288), a physician, wrote about pulmonary circulation 300 years before it was "discovered" in Europe. For Mamlūk administrative personnel, al-Qalqashandī composed an encyclopaedia in which he surveyed not only local practice but also all the information that a cultivated administrator should know. Ibn Khallikān composed one of the most important Islāmicate biographical works, a dictionary of eminent men. Sharīʿah-minded studies were elaborated: the *ʿulamāʾ* worked out a political theory that tried to make sense of the sultanate, and they also explored the possibility of enlarging on the Sharīʿah by reference to *falsafah* and Ṣūfism.

Cairo under Mamlūk rule

However, in much the same way as ash-Shāfiʿī had responded in the 9th century to what he viewed as dangerous legal diversity, another great legal and religious reformer, Ibn Taymīyah, living in Mamlūk Damascus in the late 13th and early 14th century, cautioned against such extralegal practices and pursuits. He insisted that the Sharīʿah was complete in and of itself and could be adapted to every age by any *faqih* who could analogize according to the principle of human advantage (*maṣlaḥah*). A Ḥanbalī himself, Ibn Taymīyah became as popular as his school's founder, Aḥmad ibn Ḥanbal. Like him, Ibn Taymīyah attacked all practices that undermined what he felt to be the fundamentals of Islām, including all forms of Shīʿite thought as well as aspects of Jamāʿi-Sunnite piety (often influenced by the Ṣūfīs) that stressed knowledge of God over service to him. Most visible among such practices was the revering of saints' tombs, which was condoned by the Mamlūk authorities. Ibn Taymīyah's program and popularity so threatened the Mamlūk authorities that they put him in prison, where he died. His movement did not survive, but when his ideas surfaced, in the revolutionary movement of the Wahhābīyah in the late 18th century, their lingering power became dramatically evident.

Further west, the Rūm Seljuqs at Konya submitted to the Mongols in 1243 but survived intact. They continued to cultivate the Islāmicate arts, architecture in particular. The most famous Muslim ever to live at Konya, Jalāl ad-Dīn ar-Rūmī, had emigrated from eastern Iran with his father before the arrival of the Mongols. In Konya, Jalāl ad-Dīn, attracted to Ṣūfī activities, attached himself to the master Shams ad-Dīn. The poetry inspired by Jalāl ad-Dīn's association with Shams ad-Dīn is unparalleled in Persian literature. Its recitation, along with music and movement, was a key element in the devotional activities of Jalāl ad-Dīn's followers, who came to be organized into a Ṣūfī *ṭarīqah* named the Mevleviyah (Mawlawīyah) after their title of respect for him, Mevlana ("Our Master"). In his poetry Jalāl ad-Dīn explored all varieties of metaphors, including intoxication, to describe the ineffable ecstasy of union with God.

Poetry of Jalāl ad-Dīn ar-Rūmī

Ascent of the Ottoman Turks. It was not from the Rūm Seljuqs, however, that lasting Muslim power in Anatolia was to come, but rather from one of the warrior states on the Byzantine frontier. The successive waves of Turkic migrations had driven unrelated individuals and groups across central Islāmdom into Anatolia. Avoiding the Konya state, they gravitated toward an open frontier to the west, where they began to constitute themselves, often through fictitious kinship relationships, into quasi-tribal states that depended on raiding each other and Byzantine territory and shipping. One of these, the Osmanlıs, or Ottomans, named for their founder, Osman I (ruled 1281–1324), was located not on the coast, where raiding had its limits, but in Bithynia just facing Constantinople. In 1326 they won the town of Bursa and made it their first capital. From Anatolia they crossed over into Thrace in the service of rival factions at Constantinople, then began to occupy Byzantine territory, establishing their second capital at Edirne on the European side. Their sense of legitimacy was complex. They were militantly Muslim, bound by the *ghāzī* spirit, spurred on in their intolerance of local Christians by Greek converts and traveling Ṣūfīs who gravitated to their domains. At the same time, *ʿulamāʾ* from more settled Islāmic lands to the east encouraged them to abide by the Sharīʿah and tolerate the Christians as protected non-Muslims. The Ottomans also cast themselves as deputies of the Rūm Seljuqs, who were themselves originally "deputized" by the ʿAbbāsid caliph. Finally they claimed descent from the leading Oğuz Turk families, who were natural rulers over sedentary populations. Under Murad I (ruled *c.* 1360–89) the state began to downplay its warrior fervour in favour of more conventional Islāmicate administration. Instead of relying on volunteer warriors, Murad established a regular cavalry, which he supported with land assignments, as well as a specially trained infantry force called the "new troops," Janissaries, drawn from converted captives. Expanding first through western Anatolia and Thrace, the Ottomans under Bayezid I (ruled 1389–1403) turned their eyes toward eastern and southern Anatolia; just as they had incorporated the whole, they encountered a neo-Mongol conqueror expanding into Anatolia from the east who utterly defeated their entire army in a single campaign (1402).

Timur's efforts to restore Mongol power. Timur (Tamerlane) was a Turk, not a Mongol; but he aimed to restore Mongol power. He was born a Muslim in the Syrdarya valley and served local pagan Mongol warriors and finally the Chagatai heir-apparent; but he rebelled and made himself ruler in Khwārezm in 1380. He planned to restore Mongol supremacy under a thoroughly Islāmic program. He surpassed the Mongols in terror, constructing towers out of the heads of his victims. Having established himself in Iran, he moved first on India and then on Ottoman Anatolia and Mamlūk Syria; but before he could consolidate his realm, he died. His impact was twofold: his defeat of the Ottomans inspired a comeback that would produce one of the greatest Islāmicate empires of all time, and one of the Central Asian heirs to his tradition of conquest would found another great Islāmicate empire in India. These later empires managed to find the combination of Turkic and Islāmic legitimacy that could produce the stable centralized absolutism that had eluded all previous Turkic conquerors.

ARABS

When the Fāṭimids conquered Egypt in 969, they left a governor named Zīrī in the Maghrib. By 1041 the dynasty founded by Zīrī declared its independence from the Fāṭimids, but it too was challenged by breakaways such as the Zanātah in Morocco and the Ḥammādids in Algeria. Gradually the Zīrids were restricted to the eastern Maghrib. There they were invaded from Egypt by two Bedouin Arab tribes, the Banū Halīl and the Banū Sulaym, at the instigation (1052) of the Fāṭimid ruler in Cairo. This mass migration of warriors as well as wives and children is known as the Hilālian invasion. Though initially disruptive, the Hilālian invasion had an important cultural impact; it resulted in a much greater spread of the Arabic language than had occurred in the 7th century and inaugurated the real Arabization of the Maghrib.

Hilālian invasion of the eastern Maghrib

BERBERS

When the Arab conquerors arrived in the Maghrib in the 7th century, the indigenous peoples they met were the Berbers, a group of predominantly but not entirely migratory tribes who spoke a recognizably common Hamito-Semitic language with significant dialectal variations. Berber tribes could be found from present-day Morocco to present-day Algeria, and from the Mediterranean to the Sahara. As among the Arabs, small tribal groupings of Berbers occasionally formed short-lived confederations or became involved in caravan trade. No previous conqueror had tried to assimilate the Berbers, but the Arabs quickly converted them and enlisted their aid in further conquests. Without their help, for example, Andalusia could never have been incorporated into the Islāmicate state. At first only Berbers nearer the coast were involved, but by the 11th century Muslim affiliation had begun to spread far into the Sahara.

The Ṣanhājah confederation. One particular western Saharan Berber confederation, the Ṣanhājah, was responsible for the first Berber-directed effort to control the Maghrib. The Ṣanhājah were camel herders who traded mined salt for gold with the black kingdoms of the south. By the 11th century their power in the western Sahara was being threatened by expansion both from other Berber tribes, centred at Sijilmassa, and from the Soninke state at Ghana to the south, which had actually captured their capital of Audaghost in 990. The subsequent revival of their fortunes parallels Muḥammad's revitalization of the Arabs 500 years earlier, in that Muslim ideology reinforced their efforts to unify several smaller groups. The Ṣanhājah had been in contact with Islām since the 9th century, but their distance from major centres of Muslim life had kept their knowledge of the faith minimal. In 1035, however, Yaḥyā ibn Ibrāhīm, a chief from one of their tribes, the Gudālah, went on *ḥajj.* For the Maghribi pilgrim, the cultural impact of the *ḥajj* was experienced not only in Mecca and Medina but also on the many stops along the 3,000-mile overland route. When Yaḥyā returned, he was accompanied by a teacher from Nafis (in present-day Libya), ʿAbd Allāh ibn Yasīn, who would instruct the Berbers in Islām as teachers under ʿUmar I had instructed the Arab fighters in the first Muslim garrisons. Having met with little initial success, the two are said to have retired to a *ribāṭ,* a fortified place of seclusion, perhaps as far south as an island in the Sénégal River, to pursue a purer religious life. The followers they attracted to that *ribāṭ* were known, by derivation, as *al-murābiṭūn* ("the people of the retreat"); the dynasty they founded came to be known by the same name, or Almoravids in its Anglicized form. In 1042 Ibn Yasīn declared a *jihād* against the Ṣanhājah tribes, including his own, as people who had embraced Islām but then failed to practice it properly. By his death in 1059, the Ṣanhājah confederation had been restored under an Islāmic ideology; and the conquest of Morocco, which lacked strong leadership, was under way.

Yaḥyā ibn Ibrāhīm's conversion to Islām

The Almoravid dynasty. Ibn Yasīn's spiritual role was taken by a consultative body of *ʿulamāʾ*. His successor as military commander was Abū Bakr ibn ʿUmar. While pursuing the campaign against Morocco, Abū Bakr had to go south, leaving his cousin Yūsuf ibn Tāshufīn as his deputy. When Abū Bakr tried to return, Ibn Tāshufīn turned him back to the south, where he remained until his death in 1087. Under Ibn Tāshufīn's leadership, by 1082, Almoravid control extended as far as Algiers. In 1086 Ibn Tāshufīn responded to a request for help from the Andalusian party kings, unable to defend themselves against the Christian kingdoms in the north, such as Castile. By 1110 all Muslim states in Andalusia had come under Almoravid control.

Like most other Jamāʿi-Sunnite rulers of his time, Ibn Tāshufīn had himself "appointed" deputy by the caliph in Baghdad. He also based his authority on the claim to bring correct Islām to peoples who had strayed from it. For him "correct" Islām meant the Sharīʿah as developed by the Mālikī *faqih*s, who played a key role in the Almoravid state by working out the application of the Sharīʿah to everyday problems. Like their contemporaries elsewhere, they received stipends from the government, sat in the ruler's council, went on campaign with him, and gave him recommendations (*fatwa*s) on important decisions. This was an approach to Islām far more current than the one it had replaced, but still out of touch with the liveliest intellectual developments. During the next phase of Berber activism, newer trends from the east reached the Maghrib.

Ibn Tūmart's revolt

A second major Berber movement originated in a revolt begun against Almoravid rule in 1125 by Ibn Tūmart, a settled Maṣmūdah Berber from the Atlas Mountains. Like Ibn Yasīn, Ibn Tūmart had been inspired by the *ḥajj,* which he used as an opportunity to study in Baghdad, Cairo, and Jerusalem, acquainting himself with all current schools of Islāmic thought and becoming a disciple of the ideas of the recently deceased al-Ghazālī. Emulating his social activism, Ibn Tūmart was inspired to act on the familiar Muslim dictum, "Command the good and forbid the reprehensible." His early attempts took two forms, disputations with the scholars of the Almoravid court and public chastisement of Muslims who in his view contradicted the rules of Islām; he went so far as to throw the

Almoravid ruler's sister off her horse because she was unveiled in public. His activities aroused hostility and he fled to the safety of his own people. There, like Muḥammad, he grew from teacher of a personal following to leader of a social movement.

Like many subsequent reformers, especially in Africa and other outlying Muslim lands, Ibn Tūmart used Muḥammad's career as a model. He interpreted the Prophet's rejection and retreat as an emigration (*hijrah*) that enabled him to build a community, and he divided his followers into *muhājirūn* ("fellow emigrants") and *anṣār* ("helpers"). He preached the idea of surrender to God to a people who had strayed from it. Thus could Muḥammad's ability to bring about radical change through renewal be invoked without actually claiming the prophethood that he had sealed forever. Ibn Tūmart further based his legitimacy on his claim to be a *sharīf* (descendant of Muḥammad) and the *mahdī*, not in the Shīʿite sense but in the more general sense of a human sent to restore pure faith. In his view Almoravid students of legal knowledge were so concerned with pursuing the technicalities of the law that they had lost the purifying fervour of their own founder, Ibn Yasīn. They even failed to maintain proper Muslim behaviour, be it the veiling of women in public or the condemning of the use of wine, musical instruments, and other unacceptable, if not strictly illegal, forms of pleasure. Like many Muslim revitalizers before and since, Ibn Tūmart decried the way in which the law had taken on a life of its own, and he called upon Muslims to rely on the original and only reliable sources, the Qurʾān and *ḥadīth*. Although he opposed irresponsible rationalism in the law, in matters of theological discourse he leaned toward the limited rationalism of the Ashʿarite school, which was becoming so popular in the eastern Muslim lands. Like the Ashʿarites, he viewed the unity of God as one of Islām's fundamentals and denounced any reading of the Qurʾān that led to anthropomorphism. Because he focused on attesting the unity of God (*tawḥīd*), he called his followers al-Muwaḥḥidūn (Almohads), "those who attest the unity of God." Ibn Tūmart's movement signified the degree to which Maghribis could participate in the intellectual life of Islāmdom as a whole; but his need to use Berber for his many followers who did not know Arabic also illustrates the limits of interregional discourse.

The Almohad dynasty. By 1147, 17 years after Ibn Tūmart's death, Almohads had replaced Almoravids in all their Maghribi and Andalusian territories. In Andalusia their arrival slowed the progress of the Christian Reconquista. There, as in the Maghrib, arts and letters were encouraged: an example is an important movement of *falsafah* that included Ibn Ṭufayl, Ibn al-ʿArabī, and Ibn Rushd (Latin Averroës), the Andalusian *qāḍī* and physician whose interpretations of Aristotle became so important for medieval European Christianity. During the late Almohad period in Andalusia the intercommunal nature of Islāmicate civilization became especially noticeable in the work of non-Muslim thinkers, such as Moses Maimonides, who participated in trends outside their own communities even at the expense of criticism from within. By the early 13th century, Almohad power began to decline; a defeat in 1212 at Las Navas de Tolosa by the Christian kings of the north forced a retreat to the Maghrib. But the impact of Almohad cultural patronage on Andalusia long outlasted Almohad political power; successor dynasties in surviving Muslim states were responsible for some of the highest achievements of Andalusian Muslims, among them the Alhambra palace in Granada. Furthermore, the 400-year southward movement of the Christian–Muslim frontier resulted, ironically, in some of the most intense Christian–Muslim interaction in Andalusian history. The Cid could fight for both sides; Muslims, as Mudejars, could live under Christian rule and contribute to its culture; Jews could translate Arabic and Hebrew texts into Castilian. Almohads were replaced in the Maghrib as well, through a revolt by their own governors: the Ḥafṣids in Tunis and the Marīnid Berber dynasty in Fès. There too, however, Almohad influence outlasted their political presence: both towns became centres, in distinctively Maghribi form, of Islāmicate culture and Islāmic piety.

Continued spread of Islāmic influence. As the Maghrib became firmly and distinctively Muslim, Islām moved south. The spread of Muslim identity into the Sahara and the involvement of Muslim peoples, especially the Tuareg, in trans-Saharan trade provided several natural channels of influence. By the time of the Marīnids, Ḥafṣids, and Mamlūks, several major trade routes had established criss-crossing lines of communication: from Cairo to Timbuktu, from Tripoli to Bornu and Lake Chad, from Tunis to Timbuktu at the bend of the Niger River, and from Fès and Tafilalt through major Saharan entrepôts into Ghana and Mali. The rise at Timbuktu of Mali, the first great western Sudanic empire with a Muslim ruler, attested the growing incorporation of sub-Saharan Africa into the North African orbit. The reign of Mansa Mūsā, who even went on pilgrimage, demonstrated the influence of Islām on at least the upper echelons of African society.

Travel writings of Ibn Baṭṭūṭah

The best picture of Islāmdom in the 14th century appears in the work of a remarkable Maghribi *qāḍī* and traveler, Ibn Baṭṭūṭah (1304–1368/77). In 1325, the year that Mansa Mūsā went on pilgrimage, Ibn Baṭṭūṭah also left for Mecca, from his hometown of Tangiers. He was away for almost 30 years, visiting most of Islāmdom, including Andalusia, all of the Maghrib, Mali, Syria, Arabia, Iran, India, the Maldive Islands, and, he claimed, China. He described the unity within diversity that was one of Islāmdom's most prominent features. Although local customs often seemed at variance with his notion of pure Islāmic practice, he felt at home everywhere. Despite the divisions that had occurred during Islām's 700-year history, a Muslim could attend the Friday worship session in any Muslim town in the world and feel comfortable, a claim that is difficult if not impossible to make for any other major religious tradition at any time in its history. By the time of Ibn Baṭṭūṭah's death, Islāmdom comprised the most far-flung yet interconnected set of societies in the world. As one author has pointed out, Thomas Aquinas (*c.* 1224–74) might have been read from Spain to Hungary and from Sicily to Norway; but Ibn al-ʿArabī (1165–1240) was read from Spain to Sumatra and from the Swahili coast to Kazan on the Volga River. By the end of the period of migration and renewal, Islām had begun to spread not only into sub-Saharan Africa but also into the southern seas with the establishment of a Muslim presence in the Straits of Malacca. Conversion to Islām across its newer frontiers was at first limited to a small elite, who supplemented local religious practices with Muslim ones. Islām could offer not only a unifying religious system but also social techniques, including alphabetic literacy, a legal system applicable to daily life, a set of administrative institutions, and a body of science and technology—all capable of enhancing the power of ruling elements and of tying them into a vast and lucrative trading network.

The period of migration and renewal exposed both the potentiality and the limitations of government by tribal peoples. This great problem of Islāmicate history received its most sophisticated analysis from a Maghribi Muslim named Ibn Khaldūn (1332–1406), a contemporary of Petrarch. His family had migrated from Andalusia to the Maghrib, and he himself was born in Ḥafṣid territory. He was both a *faylasūf* and a *qāḍī*, a combination more common in Andalusia and the Maghrib than anywhere else in Islāmdom. His *falsafah* was activist; he strove to use his political wisdom to the benefit of one of the actual rulers of the day. To this end he moved from one court to another before becoming disillusioned and retiring to Mamlūk Cairo as a *qāḍī*. His life thus demonstrated the importance and the constraints of royal patronage as a stimulant to intellectual creativity. In his *Muqaddimah* (the introduction to his multivolume world history) he used his training in *falsafah* to discern patterns in history. Transcending the critiques of historical method made by historians of the Būyid period, such as al-Masʿūdī, Miskawayh, and as-Suli, Ibn Khaldūn established careful standards of evidence. Whereas Muslim historians conventionally subscribed to the view that God passed sovereignty and hegemony (*dawlah*) from one dynasty to another through his divine wisdom, Ibn Khaldūn explained it in terms of a cycle of natural and inevitable stages. By his day it had become apparent that tribally organized migratory peoples, so favoured by much of the ecology of the Maghrib and the Nile-to-Oxus region, could easily acquire military superiority over settled peoples if they could capitalize on the inherently stronger group feeling (*'aṣabīyyah*) that kinship provides. Once in power, according to Ibn Khaldūn, conquering groups pass through a phase in which a small number of "builders" among them bring renewed vitality to their conquered lands. As the family disperses itself among sedentary peoples and ceases to live the hard life of migration, it becomes soft from the prosperity it has brought and begins to degenerate. Then internal rivalries and jealousies force one member of the family to become a king who must rely on mercenary troops and undermine his own prosperity by paying for them. In the end, the ruling dynasty falls prey to a new tribal group with fresh group feeling. Thus did Ibn Khaldūn call attention to the unavoidable instability of all premodern Muslim dynasties, caused by their lack of the regularized patterns of succession that were beginning to develop in European dynasties.

Consolidation and expansion (1405–1683)

After the death of Timur in 1405, power began to shift from migrating peoples to sedentary populations living in large centralized empires. After about 1683, when the last Ottoman campaign against Vienna failed, the great empires for which this period is so famous began to shrink and weaken, just as western Europeans first began to show their potential for worldwide expansion and domination. When the period began, Muslim lands had begun to recover from the devastating effects of the Black Death (1346–48), and many were prospering. Muslims had the best opportunity in history to unite the settled world, but by the end of the period, they had been replaced by Europeans as the leading contenders for this role. Muslims were now forced into direct and repeated contact with Europeans, through armed hostilities as well as through commercial interactions; and often the Europeans competed well. Yet Muslim power was so extensive, and the western Europeans such an unexpected source of competition, that Muslims were able to realize that their situation had changed only after they no longer had the strength to resist. Furthermore, the existence of several strong competitive Muslim states militated against a united response to the Europeans and could even encourage some Muslims to align themselves with the European enemies of others.

Shifting power of Muslims and Europeans

In this period, long after Islāmdom was once thought to have peaked, centralized absolutism reached its height, aided in part by the exploitation of gunpowder warfare and in part by new ways to fuse spiritual and military authority. Never before had Islāmicate ideals and institutions better demonstrated their ability to encourage political centralization, or to support a Muslim style of life where there was no organized state, be it in areas where Islām had been long established, or in areas where it was newly arrived. The major states of this period impressed contemporary Europeans; in them some of the greatest Islāmicate artistic achievements were made. In this period Muslims formed the cultural patterns that they brought into modern times, and adherence to Islām expanded to approximately its current distribution. As adherence to Islām expanded, far-flung cultural regions began to take on a life of their own. The unity of several of these regions was expressed through empire—the Ottomans in southeastern Europe, Anatolia, the eastern Maghrib, Egypt, and Syria; the Ṣafavids in Iran and Iraq; the Indo-Timurids (Mughals) in India. In these empires, Sunnite and Shī'ite became identities on a much larger scale than ever before, expressing competition between large populations; simultaneously Shī'ism acquired a permanent base from which to generate international opposition. Elsewhere, less formal and often commercial ties bound Muslims from distant locales; growing commercial and political links between Morocco and the western Sudan produced a trans-Saharan Maghribi Islām; Egyptian Islām influenced the central and eastern Sudan; and steady contacts between East Africa, South Arabia, southern Iran, southwest India, and the southern seas promoted a recognizable Indian Ocean Islām, with Persian as its lingua franca. In fact, Persian became the closest yet to an international language; but the expansion and naturalization of Islām also fostered a number of local languages into vehicles for Islāmicate administration and high culture—Ottoman, Chagatai, Swahili, Urdu, and Malay. Everywhere Muslims were confronting adherents of other religions, and new converts often practiced Islām without abandoning their previous practices. The various ways in which Muslims responded to religious syncretism and plurality continue to be elaborated to the present day.

This was a period of major realignments and expansion. The extent of Muslim presence in the Eastern Hemisphere in the early 15th century was easily discernible, but only with difficulty could one have imagined that it could soon produce three of the greatest empires in world history. From the Atlantic to the Pacific, from the Balkans to Sumatra, Muslim rulers presided over relatively small kingdoms; but nowhere could the emergence of a world-class dynasty be predicted. In Andalusia only one Muslim state, Granada, remained to resist Christian domination of the Iberian Peninsula. The Maghrib, isolated between an almost all-Christian Iberia and an eastward-looking Mamlūk Egypt and Syria, was divided between the Marīnids and Ḥafṣids. Where the Sahara shades off into the Sudanic belt, the empire of Mali at Gao was ruled by a Muslim and included several Saharan "port" cities, such as Timbuktu, that were centres of Muslim learning. On the Swahili coast, oriented as always more toward the Indian Ocean than toward its own hinterland, several small Muslim polities centred on key ports such as Kilwa. In western Anatolia and the Balkan Peninsula the Ottoman state under Sultan Mehmed I was recovering from its defeat by Timur. Iraq and western Iran were the domains of Turkic tribal dynasties known as the Black Sheep (Kara Koyunlu) and the White Sheep (Ak Koyunlu); they shared a border in Iran with myriad princelings of the Timurid line; and the neo-Mongol, neo-Timurid Uzbek state ruled in Transoxania. North of the Caspian, several Muslim khanates ruled as far north as Moscow and Kazan. In India, even though Muslims constituted a minority, they were beginning to assert their power everywhere except the south, which was ruled by Vijayanagar. In Islāmdom's far southeast, the Muslim state of Samudra held sway in Sumatra, and the rulers of the Moluccas had recently converted to Islām and begun to expand into the southern Malay Peninsula. Even where no organized state existed, as in the outer reaches of Central Asia and into southern China, scattered small Muslim communities persisted, often centred on oases. By the end of this period, Islāmdom's borders had retreated only in Russia and Iberia; but these losses were more than compensated by continuing expansion in Europe, Africa, Central Asia, and South and Southeast Asia. Almost everywhere this plethora of states had undergone realignment and consolidation, based on experimentation with forms of legitimation and structure.

Expansion of Muslim presence

OTTOMANS

Continuation of Ottoman rule. After the Ottoman state's devastating defeat by Timur, its leaders had to retain the vitality of the warrior spirit (without its unruliness and intolerance) and the validation of the Sharī'ah (without its confining independence). In 1453, Mehmed II, the Conqueror, fulfilled the warrior ideal by conquering Constantinople (soon to be known as Istanbul), putting an end to the Byzantine Empire, and subjugating the local Christian and Jewish populations. Even by then, however, a new form of legitimation was taking shape. The Ottomans continued to wage war against Christians on the frontier and to levy and convert (through the *devşirme*) young male Christians to serve in the sultan's household and army; but warriors were being pensioned off with land grants and replaced by troops more beholden to the sultan. Except for those forcibly converted, the rest of the non-Muslim population was protected for payment according to the Sharī'ah and the preference of the *ulema* (the Turkish spelling of *'ulamā'*), and organized into self-governing communities known as *millets*. Furthermore,

The conquest of Constantinople (1453)

the sultans began to claim the caliphate because they met two of its traditional qualifications: they ruled justly, in principle according to the Sharīʿah, and they defended and extended the frontiers, as in their conquest of Mamlūk Egypt, Syria, and the holy cities in 1516–17. Meanwhile they began to undercut the traditional oppositional stance of the *ulema* by building on Seljuq and Mongol practice in three ways: they promoted state-supported training of *ulema*; they defined and paid holders of religious offices as part of the military; and they aggressively asserted the validity of dynastic law alongside Sharīʿah. Simultaneously, they emphasized their inheritance of Byzantine legitimacy by transforming Byzantine symbols, such as Hagia Sophia (Church of the Divine Wisdom), into symbols for Islām; and by favouring their empire's European part, called, significantly, Rūm.

Reign of Süleyman I. The classical Ottoman system crystallized during the reign of Süleyman I, the Lawgiver (ruled 1520–66). He also pushed the empire's borders almost to their furthest limits, to the walls of Vienna in the northwest, throughout the Maghrib up to Morocco in the southwest, into Iraq to the east, and to the Yemen in the southeast. During Süleyman's reign the Ottomans even sent an expedition into the southern seas to help Aceh against the Portuguese colonizers. In theory, Süleyman presided over a balanced four-part structure: the palace household, which contained all of the sultan's wives, concubines, children, and servants; the bureaucracy (chancery and treasury); the armed forces; and the religious establishment. Important positions in the army and bureaucracy went to the cream of the *devşirme,* Christian youths converted to Islām and put through special training at the capital to be the sultan's personal "slaves." *Ulema* who acquired government posts had undergone systematic training at the major *medreses* (*madrasahs*) and so in the Ottoman state were more integrated than were their counterparts in other states; yet they were freeborn Muslims, not brought into the system as slaves of the sultan. The ruling class communicated in a language developed for their use only, Ottoman, which combined Turkic syntax with largely Arabic and Persian vocabulary. It was in this new language that so many important figures demonstrated the range and sophistication of Ottoman interests, such as the historian Mustafa Naima, the encyclopaedist Kâtip Çelebi, and the traveler Evliya Çelebi. The splendour of the Ottoman capital owed not a little to Süleyman's chief architect, the Greek *devşirme* recruit Sinan, who transformed the city's skyline with magnificent mosques and *medreses*.

Four-part administrative structure

The extent of Ottoman administration. Even in North Africa and the Fertile Crescent, where Ottoman rule was indirect, the effect of its administration, especially its land surveys and *millet* and tax systems, could be felt; remnants of the Ottoman system continue to play a role in the political life of modern states such as Israel and Lebanon, despite the fact that Ottoman control had already begun to relax by the first quarter of the 17th century. By then control of the state treasury was passing, through land grants, into the hands of local *aʿyān,* and they gradually became the real rulers, serving local rather than imperial interests. Meanwhile discontinuance of the *devşirme* and the rise of hereditary succession to imperial offices shut off new sources of vitality. Monarchs, confined to the palace during their youth, became weaker and participated less in military affairs and government councils. As early as 1630, Sultan Murad IV was presented by one of his advisers with a memorandum explaining the causes of the perceived decline and urging a restoration of the system as it had existed under Süleyman. Murad IV tried to restore Ottoman efficiency and central control, and his efforts were continued by subsequent sultans aided by a talented family of ministers known as the Köprülüs. However, during a war with Austria and Poland from 1682 to 1699, in which a major attack on Vienna failed (1683), the Ottomans suffered their first serious losses to an enemy and exposed the weakness of their system to their European neighbours. They signed two treaties, at Carlowitz in 1699 and at Passarowitz in 1718, that confirmed their losses in southeastern Europe, signified their inferiority to the Habsburg coalition, and established the defensive posture they would maintain into the 20th century.

ṢAFAVIDS

The Ṣafavid state began not from a band of *ghāzī* warriors but from a local Ṣūfī *ṭarīqah* of Ardabīl in Azerbaijan. The *ṭarīqah* was named after its founder, Shaykh Ṣafī od-Dīn (1252/53–1334), a local holy man. As for many *ṭarīqahs* and other voluntary associations, Sunnite and Shīʿite alike, affection for the family of ʿAlī was a channel for popular support. During the 15th century Shaykh Ṣafī's successors transformed their local *ṭarīqah* into an interregional movement by translating ʿAlid loyalism into full-fledged Imāmi Shīʿism. By asserting that they were the Ṣūfi "perfect men" of their time as well as descendants and representatives of the last *imām,* they strengthened the support of their Turkic tribal disciples (known as the Kizilbash, or "Red Heads," because of their symbolic 12-fold red headgear). They also attracted support outside Iran, especially in eastern Anatolia (where the anti-Ottoman Imāmi Bekṭāshī *ṭarīqah* was strong), in Syria, the Caucasus, and Transoxania. The ability of the Iranian Shīʿite state to serve as a source of widespread local opposition outside of Iran was again to become dramatically apparent many years later, with the rise of the ayatollah Ruhollah Khomeini's Islāmic Republic in the late 1970s.

Expansion in Iran and beyond. By 1501 the Ṣafavids were able to defeat the Ak Koyunlu rulers of northern Iran, whereupon their teenage leader Ismāʿīl I (ruled 1501–24) had himself proclaimed shah, using that pre-Islāmic title for the first time in almost 900 years and thereby invoking the glory of ancient Iran. The Ṣafavids thus asserted a multivalent legitimacy that flew in the face of Ottoman claims to have restored caliphal authority for all Muslims. Eventually, irritant became threat: by 1510, when Ismāʿīl had conquered all of Iran (to approximately its present frontiers) as well as the Fertile Crescent, he began pushing against the Uzbeks in the east and the Ottomans in the west, both of whom already suffered from significant Shīʿite opposition that could easily be aroused by Ṣafavid successes. Having to fight on two fronts was the most difficult military problem any Muslim empire could face. According to the persisting Mongol pattern, the army was a single force attached to the household of the ruler and moving with him at all times; so the size of an area under effective central control was limited to the farthest points that could be reached in a single campaign season. After dealing with his eastern front, Ismāʿīl turned west. At Chāldirān (1514) in northwestern Iraq, having refused to use gunpowder weapons, Ismāʿīl suffered the kind of defeat at Ottoman hands that the Ottomans had suffered from Timur. Yet through the war of words waged in a body of correspondence between Shah Ismāʿīl and the Ottoman sultan Selim I, and through the many invasions from both fronts that occurred during the next 60 years, the Ṣafavid state survived and prospered. Still living off its position at the crossroads of the trans-Asian trade that had supported all previous empires in Iraq and Iran, it was not yet undermined by the gradual emergence of more significant sea routes to the south.

Proclamation of Ismāʿīl I as shah

The first requirement for the survival of the Ṣafavid state was the conversion of its predominantly Jamāʿī-Sunnite population to Imāmi Shīʿism. This was accomplished by a government-run effort supervised by the state-appointed leader of the religious community, the *ṣadr.* Gradually forms of piety emerged that were specific to Ṣafavid Shīʿism; they centred on pilgrimage to key sites connected with the *imāms*, as well as on the annual remembering and reenacting of the key event in Shīʿite history, the caliph Yazīd I's destruction of Imām al-Ḥusayn at Karbalāʾ on the 10th of Muḥarram, AH 61 (680 CE). The 10th of Muḥarram, or ʿĀshūrāʾ, already marked throughout Islāmdom with fasting, became for Iranian Shīʿites the centre of the religious calendar. The first 10 days of Muḥarram became a period of communal mourning, during which the pious imposed suffering on themselves to identify with their martyrs of old, listened to sermons, and recited appropriate elegiac poetry. In later Ṣafavid times the name for this mourning, *taʿzīyeh,* also came to be applied to passion plays performed to reenact events surrounding al-Ḥusayn's martyrdom. Through the depths of their empathetic suffering, Shi'ites could help to overturn the injustice of al-Ḥusayn's martyrdom at the end of time, when all wrongs would be righted, all wrongdoers punished, and all true followers of the *imām*s rewarded.

Shah 'Abbās I. The state also survived because Ismā'īl's successors moved, like the Ottomans, toward a type of legitimation different from the one that had brought them to power. This development began in the reign of Ṭahmāsp (1524–76) and culminated in the reign of the greatest Ṣafavid shah, 'Abbās I (ruled 1588–1629). Since Ismā'īl's time, the tribes had begun to lose faith in the Ṣafavid monarch as spiritual leader; now 'Abbās appealed for support more as absolute monarch and less as the charismatic Ṣūfī master or incarnated *imām*. At the same time he freed himself from his unruly tribal *amīr*s by depending more and more on a paid army of converted Circassian, Georgian, and Armenian Christian captives. Meanwhile he continued to rely on a large bureaucracy headed by a chief minister with limited responsibilities; but, unlike his Ottoman contemporaries, he distanced members of the religious community from state involvement while allowing them an independent source of support in their administration of the *waqf* system. Because the Shi'ite *'ulamā'* had a tradition of independence that made them resist incorporation into the military "household" of the shah, 'Abbās' policies were probably not unpopular; but they eventually undermined his state's legitimacy. By the end of the period under discussion, it was the religious leaders, the *mujtahid*s, who would claim to be the spokesmen for the hidden *imām*. Having shared the ideals of the military patronage state, the Ottoman state became more firmly militarized and religious, as the Ṣafavid became more civilianized and secular. The long-term consequences of this breach between government and the religious institution were extensive, culminating in the establishment of the Islāmic Republic of Iran in 1978.

Ṣafavid capital at Isfahan

'Abbās expressed his new role by moving his capital in about 1597–98 to Isfahan in Fārs, the central province of the ancient pre-Islāmic Iranian empires and symbolically more Persian than Turkic. Isfahan, favoured by a high and scenic setting, became one of the most beautiful cities in the world, leading its boosters to say, "Isfahan is half the world." It came to contain, often thanks to royal patronage, myriad palaces, gardens, parks, mosques, *medrese*s, caravansaries, workshops, and public baths. Many of these still stand, including the famed Masjed-e Shah, a mosque that shares the great central mall with an enormous covered bazaar and many other structures. It was here that 'Abbās received diplomatic and commercial visits from Europeans, including a Carmelite mission from Pope Clement XIII (1604) and the adventuring Sherley brothers from Elizabethan England. Just as his visitors hoped to use him to their own advantage, 'Abbās hoped to use them to his, as sources of firearms and military technology, or as pawns in his economic warfare against the Ottomans, in which he was willing to seek help from apparently anyone, including the Russians, Portuguese, and Habsburgs.

Under Ṣafavid rule, Iran in the 16th and 17th centuries became the centre of a major cultural flowering expressed through the Persian language and through the visual arts. This flowering extended to Ṣafavid neighbour states as well—Ottomans, Uzbeks, and Indo-Timurids. Like other Shi'ite dynasties before them, the Ṣafavids encouraged the development of *falsafah* as a companion to Shi'ite esotericism and cosmology. Two major thinkers, Mīr Dāmād and his disciple Mullā Ṣadrā, members of the Ishrāqī, or illuminationist, school, explored the realm of images or symbolic imagination as a way to understand issues of human meaningfulness. The Ṣafavid period was also important for the development of Shi'ite Sharī'ah-minded studies, and it produced a major historian, Iskandar Beg Munshī, chronicler of 'Abbās' reign.

Decline of central authority. None of 'Abbās' successors was his equal, though his state, ever weaker, survived for a century. The last effective shah, Ḥusayn I (1694–1722), could defend himself neither from tribal raiding in the capital nor from interfering *mujtahid*s led by Mohammad Bāqir Majlisī (whose writings later would be important in the Islāmic Republic of Iran). In 1722, when Maḥmūd of Qandahār led an Afghan tribal raid into Iran from the east, he easily took Isfahan and destroyed what was left of central authority.

INDO-TIMURIDS (MUGHALS)

Foundation by Bābur. Although the Mongol-Timurid legacy influenced the Ottoman and Ṣafavid states, it had its most direct impact on Bābur (1483–1530), the adventurer's adventurer and founder of the third major empire of the period. Bābur's father, 'Umar Shaykh Mīrzā (died 1494) of Fergana, was one among many Timurid "princes" who continued to rule small pieces of the lands their great ancestor had conquered. After his father's death the 11-year-old Bābur, who claimed descent not only from Timur but also from Genghis Khan (on his mother's side), quickly faced one of the harshest realities of his time and place—too many princes for too few kingdoms. In his youth he dreamed of capturing Samarkand as a base for reconstructing Timur's empire. For a year after the Ṣafavid defeat of the Uzbek Muḥammad Shaybānī Khān, Bābur and his Chagatai followers did hold Samarkand, as Ṣafavid vassals; but when the Ṣafavids were in turn defeated, Bābur lost not only Samarkand but his native Fergana as well. He was forced to retreat to Kābul, which he had occupied in 1504. From there he never restored Timur's empire; rather, barred from moving north or west, he took the Timurid legacy south, to a land on which Timur had made only the slightest impression.

Victory in India

When Bābur turned toward northern India, it was ruled from Delhi by the Lodī sultans, one of many local Turkic dynasties scattered through the subcontinent. In 1526 at Pānīpat, Bābur met and defeated the much larger Lodī army. In his victory he was aided, like the Ottomans at Chāldirān, by his artillery. By his death just four years later, he had laid the foundation for a remarkable empire, known most commonly as the Mughal (*i.e.*, Mongol). It is more properly called Indo-Timurid because the Chagatai Turks were distinct from the surviving Mongols of the time and because Bābur and his successors acknowledge Timur as the founder of their power.

Bābur is also remembered for his memoirs, the *Bābur-nāmeh*. Written in Chagatai, then an emerging Islāmicate literary language, his work gives a lively and compelling account of the wide range of interests, tastes, and sensibilities that made him so much a counterpart of his contemporary, the Italian Niccolò Machiavelli (1469–1527).

Reign of Akbar. Süleyman's and 'Abbās' counterpart in the Indo-Timurid dynasty was their contemporary, Akbar (ruled 1556–1605), the grandson of Bābur. At the time of his death, he ruled all of present-day India north of the Deccan Plateau and Gondwana, and more: one diagonal of his empire extended from the Hindu Kush to the Bay of Bengal; the other, from the Himalayas to the Arabian Sea. Like its contemporaries to the west, particularly the Ottomans, this state endured because of a regularized and equitable tax system that provided the central treasury with funds to support the ruler's extensive building projects as well as his *manṣabdār*s, the military and bureaucratic officers of the imperial service. For these key servants, Akbar, again like his counterparts to the west, relied largely on foreigners who were trained especially for his service. Like the Janissaries, the *manṣabdār*s were not supposed to inherit their offices, and, although they were assigned lands to supervise, they themselves were paid through the central treasury to assure their loyalty to the interests of the ruler.

Akbar's centralized state

Although Akbar's empire was, like Süleyman's and 'Abbās', a variation on the theme of the military patronage state, his situation, and consequently many of his problems, differed from theirs in important ways. Islām was much more recently established in most of his empire than in either of the other two, and Muslims were not in the majority. Although the other two states were not religiously or ethnically homogeneous, the extent of their internal diversity could not compare with Akbar's, where Muslims and non-Muslims of every stripe alternately coexisted and came into conflict—Jacobites (members of the Monophysite Syrian church), Ṣūfīs, Ismāʿīlī Shīʿites, Zoroastrians, Jains, Jesuits, Jews, and Hindus. Consequently, Akbar was forced even more than the Ottomans to confront and address the issue of religious plurality. The option of aggressive conversion was virtually impossible in such a vast area, as was any version of the Ottoman *millet* system in a setting in which hundreds if not thousands of *millets* could be defined.

In some ways Akbar faced, in exaggerated form, the situation that the Arab Muslims faced when they were a minority in the Nile-to-Oxus region in the 7th–9th century. Granting protected status to non-Muslims, even those who were not really "peoples of the book" in the original sense, with an organized religion of their own, was legally and administratively justifiable; but unless they could be kept from interacting too much with the Muslim population, Islām itself could be affected. The power of Ṣūfī *ṭarīqahs* like the influential Chishtis, and of the Hindu mystical movement of Gurū Nānak, were already promoting intercommunal interaction and cross-fertilization. Akbar's response was different from that of the ʿAbbāsid caliph al-Mahdi. Instead of institutionalizing intolerance of non-Muslim influences, and instead of hardening communal lines, Akbar banned intolerance and even the special tax on non-Muslims. To keep the *ʿulamāʾ* from objecting, he tried, for different reasons than had the Ottomans and Ṣafavids, to tie them to the state financially. His personal curiosity about other religions was exemplary; with the help of Abuʾl-Faẓl, his Ṣūfī adviser and biographer, he established a kind of salon for religious discussion. A very small circle of personal disciples seems to have emulated Akbar's own brand of *tawḥīd-i ilāhī* ("divine oneness"). This appears to have been a general monotheism akin to what the *ḥanīfs* of Mecca, and Muḥammad himself, had once practiced, as well as to the boundary-breaking pantheistic awareness of great Ṣūfīs like ar-Rūmī and Ibn al-ʿArabī, who was very popular in South and Southeast Asia. Akbar combined toleration for all religions with condemnation of practices that seemed to him humanly objectionable, such as enslavement and the immolation of widows.

Continuation of the empire. For half a century, Akbar's first two successors, Jahāngīr and Shāh Jahān, continued his policies. A rebuilt capital at Delhi was added to the old capitals of Fatehpur Sīkri and Āgra, site of Shāh Jahān's most famous building, the Tāj Mahal. The mingling of Hindu and Muslim traditions was expressed in all the arts, especially in naturalistic and sensuous painting; extremely refined and sophisticated design in ceramics, inlay-work, and textiles; and in delicate yet monumental architecture. Shāh Jahān's son, Dārā Shikōh (1615–59), was a Ṣūfī thinker and writer who tried to establish a common ground for Muslims and Hindus. In response to such attempts, a Sharīʿah-minded movement of strict communalism arose, connected with a leader of the Naqshbandī *ṭarīqah* named Shaykh Aḥmad Sirhindī. With the accession of Aurangzeb (ruled 1659–1707) the tradition of ardent ecumenicism, which would reemerge several centuries later in a non-Muslim named Mohandas K. (Mahatma) Gandhi, was replaced with a stricter communalism that imposed penalties on protected non-Muslims and stressed the shah's role as leader of the Muslim community, by virtue of his enforcing the Sharīʿah. Unlike the Ottoman and Ṣafavid domains, the Indo-Timurid empire was still expanding right up to the beginning of the 18th century; but the empire began to disintegrate shortly after the end of Aurangzeb's reign, when Ṣafavid and Ottoman power were also declining rapidly.

Rise of communalism

Between the 15th and 18th century the use of coffee, tea, and tobacco, despite the objections of the *ʿulamāʾ*, became common in all three empires. Teahouses became important new centres for male socializing, in addition to the home, the mosque, the marketplace, and the public bath. (Female socializing was restricted largely to the home and the bath.) In the teahouses men could practice the already well-developed art of storytelling and take delight in the clever use of language. *The Thousand and One Nights* (*Alf laylah wa laylah*), the earliest extant manuscripts of which date from this period, and the stories of the Arabian hero ʿAntar must have been popular, as were the tales of a wise fool known as Mullah Nasroddin in Persian (Nasreddin), Hoca in Turkish, and Juḥā in Arabic. The exploits of Nasroddin, sometimes in the guise of Ṣūfī dervish or royal adviser, often humorously portray centralized absolutism and mysticism: "Nasroddin was sent by the King to investigate the lore of various kinds of Eastern mystical teachers. They all recounted to him tales of the miracles and the sayings of the founders and great teachers, all long dead, of their schools. When he returned home, he submitted his report, which contained the single word 'Carrots.' He was called upon to explain himself. Nasroddin told the King: 'The best part is buried; few know—except the farmer—by the green that there is orange underground; if you don't work for it, it will deteriorate; there are a great many donkeys associated with it.' "

TRANS-SAHARAN ISLĀM

When the Ottomans expanded through the southern Mediterranean coast in the early 16th century, they were unable to incorporate Morocco, where a new state had been formed in reaction to the appearance of the Portuguese. The Portuguese were riding the momentum generated by their own seaborne expansion as well as by the fulfillment of the Reconquista and the establishment of an aggressively intolerant Christian regime in the centre of the Iberian Peninsula. In Morocco, it was neither the fervour of warriors nor Shīʿite solidarity nor Timurid restoration that motivated the formation of a state; rather it was a very old form of legitimacy that had proved to be especially powerful in Africa, that of the *sharīfs*, descendants of Muḥammad. It had last been relied on with the Idrīsids; now the *sharīfs* were often associated with Ṣūfī holy men, known as marabouts. It was one such Ṣūfī, Sīdī Barakāt, who legitimated the Saʿdī family of *sharīfs* as leaders of a *jihad* that expelled the Portuguese and established an independent state (1511–1603) strong enough to expand far to the south. Meanwhile the greatest Muslim kingdom of the Sudan, Songhai, was expanding northward; and its growing control of major trade routes into Morocco provoked Moroccan interference. Invaded in 1591, Songhai was ruled as a Moroccan vassal for 40 years, during which time Morocco itself was experiencing political confusion and instability. Morocco was reunited in 1668 by the ʿAlawite *sharīfs*. A holy family of Sijilmassa, they were brought to power by Arab tribal support, which they eventually had to replace with a costly army of black slaves. Like the Saʿdīs, they were legitimated in two ways: by the recognition of leading Ṣūfīs and by the special spiritual quality (*barakah*) presumed to have passed to them by virtue of their descent from the Prophet through ʿAlī. Although they were not Shīʿites, they cultivated charismatic leadership that undermined the power of the *ʿulamāʾ* to use the Sharīʿah against them. They also recognized the limits of their authority as absolute monarchs, dividing their realm into the area of authority and the area of no authority (where many of the Berber tribes lived). Thus the Moroccan *sharīfs* solved the universal problems of legitimacy, loyalty, and control in a way tailored to their own situation.

Islāmic states in the Sudanic region

While the Saʿdī dynasty was ruling in Morocco, but long before its incursions into the Sahara, a number of small Islāmic states were strung from one end of the Sudanic region to the other: Senegambia, Songhai, Aïr, Mossi, Nupe, Hausa, Kanem-Bornu, Darfur, and Funj. Islām had come to these areas along trade and pilgrimage routes, especially through the efforts of a number of learned teaching-trading families such as the Kunta. Ordinarily the ruling elites became Muslim first, employing the skills of Arab immigrants, traders, or travelers, and taking political and commercial advantage of the Arabic language and the Sharīʿah without displacing indigenous religious practices or legitimating principles. By the 16th century the Muslim states of the Sudanic belt were in contact not only with the major Muslim centres of the Maghrib and Egypt, but also with each other through an emerging trans-Sudanic pilgrimage route. Furthermore, Islām had by then become well enough established to provoke efforts at purification comparable with the Almoravid movement of the 11th century. Sometimes these efforts were gradualist and primarily educational, as was the case with the enormously influential Egyptian scholar as-Suyūṭī (1445–1505). His works, read by many West African Muslims for centuries after his death, dealt with numerous subjects, including the coming of the *mahdī* to restore justice and strengthen Islām. He also wrote letters to Muslim scholars and rulers in West Africa more than 2,000 miles away, explaining the Sharīʿah and encouraging its careful observance.

Other efforts to improve the observance of Islām were more militant. Rulers might forcibly insist on an end to certain non-Muslim practices, as did Muḥammad Rumfa (ruled 1463–99) in the Hausa city-state of Kano, or Muḥammad I Askia, the greatest ruler of Songhai (ruled 1493–1528). Often, as in the case of both of these rulers, militance was encouraged by an aggressive reformist scholar like al-Maghīlī (flourished 1492), whose writings detailed the conditions that would justify a *jihād* against Muslims who practiced their faith inadequately. Like many reformers, al-Maghīlī identified himself as a *mujaddid,* a figure expected to appear around the turn of each Muslim century. (The 10th century of the *hijrah* era began in 1494.) To the east in Ethiopia, an actual *jihād* was carried out by Aḥmad Grāñ (*c.* 1506–43), in the name of opposition to the Christian regime and purification of "compromised" Islām. Further to the east, a conquest of Christian Nubia by Arab tribes of Upper Egypt resulted in the conversion of the pagan Funj to Islām and the creation of a major Muslim kingdom there. Although most indigenous West African scholars looked to foreigners for inspiration, a few began to chart their own course. In Timbuktu, where a rich array of Muslim learning was available, one local scholar and member of a Tukulor learned family, Aḥmad Bābā, was writing works that were of interest to North African Muslims. Local histories written in Arabic also survive, such as the *Taʾrīkh al-fattāsh* (written by several generations of the Kāti family, from 1519 to 1665), a chronological history of Songhai, or as-Saʿdī's *Taʾrīkh al-Sūdān* (completed in 1655). By the end of the period of consolidation and expansion, Muslims in the Sudanic belt were being steadily influenced by North African Islām but were also developing distinctive traditions of their own.

INDIAN OCEAN ISLĀM

A similar relationship was simultaneously developing across another "sea," the Indian Ocean, which tied South and Southeast Asian Muslims to East African and south Arabian Muslims the way the Sahara linked North African and Sudanic Muslims. Several similarities are clear: the alternation of advance and retreat, the movement of outside influences along trade routes, and the emergence of significant local scholarship. There were differences, too: Indian Ocean Muslims had to cope with the Portuguese threat and to face Hindus and Buddhists more than pagans, so that Islām had to struggle against sophisticated and refined religious traditions that possessed written literature and considerable political power.

The Aceh state

The first major Muslim state in Southeast Asia, Aceh, was established around 1524 in northern and western Sumatra in response to more than a decade of Portuguese advance. Under Sultan Iskandar Muda (ruled 1608–37), Aceh reached the height of its prosperity and importance in the Indian Ocean trade, encouraging Muslim learning and expanding Muslim adherence. By the end of the 17th century, Aceh's Muslims were in touch with major intellectual centres to the west, particularly in India and Arabia, just as West African Muslims were tied to centres across the Sahara. Because they could draw on many sources, often filtered through India, Sumatran Muslims may have been exposed to a wider corpus of Muslim learning than Muslims in many parts of the heartland. Aceh's scholarly disputes over Ibn al-ʿArabī were even significant enough to attract the attention of a leading Medinan, Ibrahim al-Kurani, who in 1640 wrote a response. The same kind of naturalization and indigenization of Islām that was taking place in Africa was also taking place here; for example, ʿAbd ar-Raʾūf of Singkel, after studying in Arabia from about 1640 to 1661, returned home, where he made the first "translation" of the Qurʾān into Malay, a language that was much enriched during this period by Arabic script and vocabulary. This phenomenon extended even to China. Liu Chih, a scholar born around 1650 in Nanking, created serious Islāmicate literature in Chinese, including works of philosophy and law.

In the early 17th century another Muslim commercial power emerged when its ruler, the prince of Tallo, converted; Macassar (now Ujung Pandang) became an active centre for Muslim competition with the Dutch into the third quarter of the 17th century, when its greatest monarch, Ḥasan ad-Dīn (ruled 1631–70), was forced to cede his independence. Meanwhile, however, a serious Islāmic presence was developing in Java, inland as well as on the coasts; by the early 17th century the first inland Muslim state in Southeast Asia, Mataram, was established. There Ṣūfī holy men performed a missionary function similar to that being performed in Africa. Unlike the more seriously Islāmized states in Sumatra, Mataram suffered, as did its counterparts in West Africa, from its inability to suppress indigenous beliefs to the satisfaction of the more conservative *ʿulamāʾ*. Javanese Muslims, unlike those in Sumatra, would have to struggle for centuries to negotiate the confrontation between Hindu and Muslim cultures. Their situation underscores a major theme of Islāmicate history through the period of consolidation and expansion; that is, the repeatedly demonstrated absorptive capacity of Muslim societies, a capacity that was soon to be challenged in unprecedented ways.

Reform, dependency, and recovery (1683 to the present)

The history of the Muslims in modern times has often been explained in terms of the impact of "the West." From this perspective, the 18th century was a period of degeneration and a prelude to European domination, symbolized by Napoleon's conquest of Egypt in 1798. Given the events of the 1980s, however, it is possible to argue that the period of Western domination was an interlude in the ongoing development of indigenous styles of modernization. In order to examine that hypothesis, it is necessary to begin the "modern" period with the 18th century, when activism and revival were present throughout Islāmdom. The three major Muslim empires did experience a decline during the 18th century, as compared to their own earlier power and to the rising powers in Europe; but most Muslims were not yet aware that Europe was partly to blame. Similar decline had occurred many times before, a product of the inevitable weaknesses of the military conquest state turned into centralized absolutism, overdependence on continuous expansion, weakening of training for rule, the difficulty of maintaining efficiency and loyalty in a large, complex royal household and army, and the difficulty of maintaining sufficient revenues for an increasingly lavish court life. Furthermore, population increased, as it did almost everywhere in the 18th-century world, just as inflation and expensive reform reduced income to central governments. Given the insights of Ibn Khaldūn, however, one might have expected a new group with a fresh sense of cohesiveness to restore political strength.

Economic and social strengths in western Europe

Had Muslims remained on a par with all other societies, they might have revived. But by the 18th century one particular set of societies in western Europe had developed an economic and social system capable of transcending the 5,000-year-old limitations of the agrarian-based settled world as defined by the Greeks (who called it Oikoumene). Unlike most of the lands of Islāmdom, those societies were rich in natural resources (especially the fossil fuels that could supplement human and animal power) and poor in space for expansion. Cut off by Muslims from controlling land routes from the East, European explorers had built on and surpassed Muslim seafaring technology to compete in the southern seas and discover new sea routes—and, accidentally, a new source of wealth in the Americas. In Europe, centralized absolutism, though an ideal, had not been the success it was in Islāmdom. Emerging from the landed classes rather than from the cities, it had benefited from and been constrained by independent urban commercial classes. In Islāmdom, the power of merchants had been inhibited by imperial overtaxation of local private enterprise, appropriation of the benefits of trade, and the privileging of foreign traders through agreements known as the Capitulations.

In Europe independent financial and social resources promoted an unusual freedom for technological experimentation and, consequently, the technicalization of other areas of society as well. Unlike previous innovations in the Oikoumene, Europe's technology could not easily be diffused to societies that had not undergone the prerequisite fundamental social and economic changes. Outside of Europe, gradual assimilation of the "new," which had characterized change and cultural diffusion for 5,000 years, had to be replaced by hurried imitation, which proved enormously disorienting. This combination of innovation and imitation produced an unprecedented and persisting imbalance among various parts of the Oikoumene. Muslims' responses paralleled those of other "non-Western" peoples but were often filtered through and expressed in peculiarly Islāmic or Islāmicate symbols and motifs. The power of Islām as a source of public values had already waxed and waned many times; it intensified in the 18th and 19th centuries, receded in the early 20th century, and surged again after the mid-20th century. Thus European colonizers appeared in the midst of an ongoing process that they greatly affected but did not completely transform.

PRE-COLONIAL REFORM AND EXPERIMENTATION (1683–1818)

From the mid-17th century through the 18th and early 19th centuries certain Muslims expressed an awareness of internal weakness. In some areas, Muslims were largely unaware of the rise of Europe; in others, such as India, Sumatra, and Java, the 18th century actually brought European control. Responses to decline, sometimes official and sometimes unofficial, sometimes Islāmizing, sometimes Europeanizing, fell into two categories, as the following examples demonstrate.

In some areas, leaders attempted to revive existing political systems. In Iran, for example, attempts at restoration combined military and religious reform. Around 1730 a Turk from Khorāsān named Nāder Qolī Beg reorganized the Ṣafavid army in the name of the Ṣafavid shah, whom he replaced with himself in 1736. Nāder Shāh extended the borders of the Ṣafavid state further than ever; he even defeated the Ottomans and may have been aspiring to be the leader of all Muslims. To this end he made overtures to neighbouring rulers, seeking their recognition by trying to represent Iranian Shīʿism as a *madhhab* alongside the Sunnite *madhhabs*. After he was killed in 1747, however, his reforms did not survive and his house disintegrated. Karīm Khān Zand, a general from Shīrāz, ruled in the name of the Ṣafavids but did not restore real power to the shah. By the time the Qājārs (1779–1925) managed to resecure Iran's borders, reviving Ṣafavid legitimacy was impossible.

Restoration in the Ottoman Empire

In the Ottoman Empire, restoration involved selective imitation of things European. Its first phase, from 1718 to 1730, is known as the Tulip Period, because of the cultivation by the wealthy of a Perso-Turkish flower then popular in Europe. Experimentation with European manners and tastes was matched by experimentation with European military technology. Restoration depended on reinvigorating the military, the key to earlier Ottoman success, and Christian Europeans were hired for the task. After Nāder Shāh's defeat of the Ottoman army, this first phase of absolutist restoration ended, but the pursuit of European fashion had become a permanent element in Ottoman life. Meanwhile, central power continued to weaken, especially in the area of international commerce. The certificates of protection that had accompanied the Capitulations arrangements for foreign nationals were extended to non-Muslim Ottoman subjects, who gradually oriented themselves toward their foreign associates. The integration of such groups into the Ottoman state was further weakened by the recognition, in the disastrous Treaty of Küçük Kaynarca (1774), of the Russian tsar as protector of the Ottoman's Greek Orthodox *millet*.

A second stage of absolutist restoration occurred under Selim III, who became sultan in the first year of the French Revolution and ruled until 1807. His military and political reforms, referred to as the New Order (Nizam-ı Cedid), went beyond the Tulip Period in making use of things European; for example, the enlightened monarch, as exemplified by Napoleon himself, became an Ottoman ideal. Here, as in Egypt under Muḥammad ʿAlī (reigned 1805–48), the famed core of Janissaries that had been a source of Ottoman strength was destroyed and replaced with European-trained troops.

In other areas, leaders envisioned or created new social orders that were self-consciously Islāmic. The growing popularity of westernization and a decreasing reliance on Islām as a source of public values was counterbalanced in many parts of Islāmdom by all sorts of Islāmic activism, ranging from educational reform to *jihād*. "Islāmic" politics often were marked by an oppositional quality that drew on long-standing traditions of skepticism about government. Ṣūfism could play very different roles. In the form of renovated *ṭarīqahs* it could support reform and stimulate pan-Islāmic awareness. Ṣūfīs often encouraged the study of *ḥadīth* so as to establish the Prophet Muḥammad as a model for spiritual and moral reconstruction and to invalidate many unacceptable traditional or customary Islāmic practices. Ṣūfī *ṭarīqahs* provided interregional communication and contact and an indigenous form of social organization that could even lead to the founding of a dynasty, as in the case of the Libyan monarchy.

Ṣūfism could also be condemned as a source of degeneracy. The most famous and influential militant anti-Ṣūfī movement arose in the Arabian Peninsula and called itself al-Muwaḥḥidūn ("the Monotheists"); but it came to be known as Wahhābīyah, after its founder, Muḥammad ibn ʿAbd al-Wahhāb (1703–92). Inspired by Ibn Taymīyah (see above *Migration and renewal (1041–1405)*), Ibn al-Wahhāb argued that the Qurʾān and *sunnah* could provide the basis for a reconstruction of Islāmic society out of the degenerate form in which it had come to be practiced. Islām itself was not an inhibiting force; "traditional" Islām was. Far from advocating the traditional, the Wahhābīs argued that what had become traditional had strayed very far from the fundamental, which can always be found in the Qurʾān and *sunnah*. The traditional they associated with blind imitation (*taqlīd*); reform, with making the pious personal effort (*ijtihād*) necessary to understand the fundamentals. Within an Islāmic context, this type of movement was not conservative, because it sought not to conserve what had been passed down but to renew what had been abandoned. The Wahhābī movement attracted the support of a tribe in the Najd led by Muḥammad ibn Saʿūd. Although the first state produced by this alliance did not last, it laid the foundations for the existing Saudi state in Arabia and inspired similar activism elsewhere down to the present day.

Islāmic activism in West Africa

In West Africa a series of activist movements appeared from the 18th century into the 19th. There as in Arabia, Islāmic activism was directed less at non-Muslims than at Muslims who had gone astray. As in many of Islāmdom's outlying areas, emergent groups of indigenous educated, observant Muslims, such as the Tukulor, were finding the casual, syncretistic, opportunistic nature of official Islām to be increasingly intolerable. Such Muslims were inspired by reformist scholars from numerous times and places—al-Ghazālī, as-Suyūṭī, Maghili; by a theory of *jihād* comparable to that of the Wahhābīs; and by expectations of a *mujaddid* as the Islāmic century turned in AH 1200 (AD 1785). In what is now northern Nigeria, the discontent of the 1780s and '90s erupted in 1804, when Usman dan Fodio declared a *jihād* against the Hausa rulers. Others followed, among them Muhammad al-Jaylani in Aïr, Shehuh Ahmadu Lobbo in Macina, al-Ḥajj ʿUmar Tal (a member of the reformist Tijānī *ṭarīqah*) in Fouta Djallon, and Samory in the Malinke (Mandingo) states. *Jihād* activity continued for a century; it again became millennial near the turn of the next Muslim century in AH 1300 (AD 1882), as the need to resist against European occupation became more urgent. For example, Muḥammad Aḥmad declared himself to be the *mahdī* in the Sudan in 1881.

In the Indian Ocean area, Islāmic activism was more often intellectual and educational. Its best exemplar was Shāh Walī Allāh of Delhi (1702–62), the spiritual ancestor of many later Indian Muslim reform movements. During his lifetime the collapse of Muslim political power was painfully evident. He tried to unite the Muslims of India, not around Ṣūfism as Akbar had tried to do, but around the Sharīʿah. Like Ibn Taymīyah, he understood the Sharīʿah to be based on firm sources—Qurʾān and *sunnah*—that could with pious effort be applied to present circumstances. Once again, the study of *ḥadīth* provided a rich array of precedents and inspired a positive spirit of social reconstruction akin to that of the Prophet Muḥammad.

DEPENDENCY (1818–1962)

The many efforts to revive and resist were largely unsuccessful. By 1818, British hegemony over India was complete; and many other colonies and mandates followed between then and the aftermath of World War I. Not all Muslim territories were colonized, but nearly all experienced some kind of dependency, be it psychological, political, technological, cultural, or economic. Perhaps only the Saudi regime in the central parts of the Arabian Peninsula could be said to have escaped any kind of dependency; but even there oil exploration, begun in the 1930s, brought European interference. In the 19th century westernization and Islāmic activism coexisted and competed. By the turn of the 20th century secular ethnic nationalism had become the most common mode of protest in Islāmdom; but the spirit of Islāmic reconstruction was also kept alive, either in conjunction with secular nationalism or in opposition to it.

In the 19th-century Ottoman Empire, selective westernization coexisted with a reconsideration of Islām. The program of reform known as the Tanzimat, which was in effect from 1839 to 1876, aimed to emulate European law and administration by giving all Ottoman subjects, regardless of religious confession, equal legal standing and by limiting the powers of the monarch. In the 1860s a group known as the Young Ottomans tried to identify the basic principles of European liberalism and even love of nation with Islām itself. In Iran, the Qājār shahs brought in a special "Cossack Brigade," trained and led by Russians, while at the same time the Shīʿite *mujtahid*s viewed the decisions of their spiritual leader as binding on all Iranian Shīʿites and declared themselves to be independent of the shah. (One Shīʿite revolt, that of the Bāb [died 1850], led to a whole new religion, Bahāʾī.) Like the Young Ottomans, Shīʿite religious leaders came to identify with constitutionalism in opposition to the ruler.

Reaction against Europeans and westernization

Islāmic protest often took the form of *jihād* against the Europeans: by Southeast Asians against the Dutch; by the Sanūsī *ṭarīqah* over Italian control in Libya; by the Mahdist movement in the Sudan; or by the Ṣaliḥī *ṭarīqah* in Somalia, led by Sayyid Muḥammad ibn ʿAbd Allāh Ḥasan, who was tellingly nicknamed the Mad Mullah by Europeans. Sometimes religious leaders, like those of the Shīʿites in Iran, took part in constitutional revolutions (1905–11). Underlying much of this activity was a pan-Islāmic sentiment that drew on very old conceptions of the *ummah* as the ultimate solidarity group for Muslims. Three of the most prominent Islāmic reconstructionists were Jamāl ad-Dīn al-Afghānī, his Egyptian disciple Muḥammad ʿAbduh, and the Indian poet Sir Muḥammad Iqbāl. All warned against blind pursuit of Westernization, arguing that the blame for the weaknesses of Muslims lay not with Islām, but rather with Muslims themselves, because they had lost touch with the progressive spirit of social, moral, and intellectual reconstruction that had made early Islāmicate civilization one of the greatest in human history. Although al-Afghānī, who taught and preached in many parts of Islāmdom, acknowledged that organization by nationality might be necessary, he viewed it as inferior to Muslim identity. He further argued that Western technology could advance Muslims only if they retained and cultivated their own spiritual and cultural heritage. He pointed out that at one time Muslims had been intellectual and scientific leaders in the world, identifying a Golden Age under the ʿAbbāsid caliphate and pointing to the many contributions Muslims had made to "the West." Like al-Afghānī, Iqbāl assumed that without Islām Muslims could never regain the strength they had possessed when they were a vital force in the world, united in a single international community and unaffected by differences of language or ethnos. This aggressive recovery of the past became a permanent theme of Islāmic reconstruction. In many regions of Islāmdom the movement known as Salafiyah also identified with an ideal time in history, that of the "pious ancestors" (*salaf*) in the early Muslim state of Muḥammad and his companions, and advocated past-oriented change to bring present-day Muslims up to the progressive standards of an earlier ideal. In addition to clearly Islāmic thinkers, there were others, such as the Egyptian Muṣṭafā Kāmil, whose nationalism was not simply secular. Kāmil saw Egypt as simultaneously European, Ottoman, and Muslim. The Young Turk Revolution of 1908 was followed by a period in which similarly complex views of national identity were discussed in the Ottoman Empire.

RECOVERY (1922 TO THE PRESENT)

Progress of secular nationalism. Despite the ideological appeal of such positions, the need to throw off European control promoted the fortunes of secular nationalism and other narrower forms of loyalty. Especially after Japan's defeat of Russia in 1905, nationalist fervour increased. Sometimes it was associated with related ideologies, such as pan-Arabism, pan-Turkism, or Arab socialism. Many nationalists enthusiastically admired things European despite the fact that they were committed to resisting or removing European control. Often accepting European assessments of traditional religion as a barrier to modernization, many nationalists sought an identity in the pre-Islāmic past. Kemal Atatürk looked to the Turkic past in Central Asia and Anatolia to transform Ottomanism into a Turkish identity not dependent on Islām. "Islāmic" dress was discouraged. Muslim males, who prayed with covered heads, were now asked to replace the fez, which could be kept on during prayer, with the brimmed hat, which could not. Arabic script, too closely associated with Islām, was replaced with the Roman, after the Cyrillic (the alphabet of Central Asian Turks) had been considered and rejected. In Iran, Reza Shah Pahlavi argued that the Islāmic period was but an accidental interlude in the continuous history, since Achaemenid times, of Iran as a unified entity. The Egyptian Taha Hussein connected his country's national identity with Pharaonic times and with Mediterranean–European culture; and therefore it could easily partake of modern Western civilization. Christians were thus as much Egyptians as were Muslims; the accompanying development of a standard literary Arabic, *fuṣḥā*, emphasized the unity of all Arabs, regardless of confession. These approaches allowed, indeed required, all religious communities to partake of a single legal and societal system, at the price of denying the public relevance of a primary loyalty for the majority of the population.

Other nationalists made more of Islām. In Saudi Arabia and Pakistan, for example, Islām played a primary role in the formation of a national identity. In Pakistan it provided, according to the statesman Mohammed Ali Jinnah, an alternative for Muslims who would otherwise have to share in an identity defined by a Hindu majority. In many Arab countries, especially in the Maghrib, secular nationalism's downgrading of Islām was muted by a qualified acceptance of Islām as one, but not the only, important source of loyalty. At the same time there were Muslims who opposed nationalism altogether. In India, Mawlanā Abuʾl-ʿAlāʾ Mawdūdī, who was the founder of the Jamāʿat-i Islāmī, opposed both secular and religious nationalism and argued for the Islāmization of society and an Islāmic alternative to nationalism. In Egypt, Sayyid Quṭb and Ḥasan al-Bannāʾ, who were the mentors of the Muslim Brotherhood, fought for the educational, moral, and social reform of an Islāmic Egypt and indeed of all Islāmdom.

Opposition to nationalism

Creating national identities. Only a few existing states where Muslims predominate, such as Turkey and Saudi Arabia, had no colonial interval; most became independent after World War II. An even larger number of countries have Muslim minorities. Like the citizens of many new nations, Muslims have not found the creation of national identities to be easy, especially considering the pace at which it has had to occur. More than two-thirds of the world's nations have come into existence since the end of World War II; foreign dependency is a living memory for many of their citizens, or at least for the parents and grandparents of their citizens. Many of them are not nation-states—that is, states established by a group of people who decided that they belonged together and therefore went about acquiring sovereignty over a territory—but rather are state-nations, composed of groups of people who acquired or were given sovereignty over a territory and then had to develop a sense of nationality. The most obvious state-nations are Syria, Iraq, Lebanon, and Jordan. All resulted from the interaction of intra-European rivalry and diplomacy with the aspirations of a prominent Ottoman-Hāshimite sharifian family in Mecca to create a single Arab state in the East. Instead of a single state, however, three monarchies emerged: the kingdom of Ḥusayn ibn ʿAlī in the Hejaz (to be replaced by the Saudis), the kingdom of Fayṣal I in Iraq (because he had to be compensated for being ousted from Syria), and the kingdom of Abdullah in Transjordan. Lebanon was carved from French Syria with borders that would establish a bare Christian majority loyal to the French. In Ottoman Palestine, Jewish nationalists clashed with Arab nationalists at a time when both groups felt betrayed by the British. In subsequent armed clashes, Zionist groups defended a set of boundaries as artificial as many others, creating a state that has remained a target for anti-imperialist sentiment. Eventually, Jewish nationalism spawned another nationalism, that of the Palestinians, inchoate before the founding of Israel but crystallized by the failure of any party to the conflict—Arab states, foreign powers, Palestinian leaders, or Israel itself—to make a place for most of the former Arab residents of Palestine. However, in September 1993 a peace agreement between the Palestine Liberation Organization and Israel established Palestinian self-rule in Israeli-occupied territory, although implementation proved to be arduous.

Many Muslim countries were united by negative nationalism, aimed at ejecting a common enemy; but turning negative into positive has been difficult. Rarely have the groups that achieved independence survived. Often, as in Libya or Iraq or Egypt, further revolutions have occurred, in many cases led by the military, whose role as a vehicle for modernization cannot be underestimated. Subsequent governments have had to deal with the social and economic problems that plague all developing countries, as well as with regional rivalries and conflicts. Almost nowhere did the colonizers leave an infrastructure sufficient to support the growth of population that European medicine and hygiene had produced.

Relation of religion and nationality. Given the multi-communal structure of premodern Muslim societies, the relation between religion and nationality has been another major problem. Nationalism has frequently led to competition and rivalry among a new nation's religious communities. As they became independent, citizens of the nations of Islāmdom could draw on no direct equivalent of national identity. The broadest identity was provided by membership in a pan-territorial community like the *ummah* of all Muslims or the Greek Orthodox Church or the Turkic tribes; the narrowest, family or neighbourhood. In the middle of the spectrum was membership in a local confessional community, with all its implications of status, occupation, manners, and customs. Citizens of the new nations would theoretically have to find an identity that could subsume and supersede all others; and the rulers of new nations would have to take the unprecedented step of declaring all citizens subject to the same law, rather than members of quasi-autonomous, self-governing religious communities with their own legal systems. Yet the significance of being a member of a religious community could not easily be undone or replaced.

Rivalry among religious communities

Many countries inherited a relatively simple form of this problem: the people within their borders were primarily of one faith, Islām, and of one form of that faith, the Sunnite. That majority adherence could in some way be associated with or bolster the national identity, while discomfiting only a small number of people. Turkey, Iran, Jordan, Indonesia, Yemen, and all the states of North Africa and the Arabian Peninsula fall into this category. Even so, religious minorities in these countries (such as the Armenians) suffered and shrank; for Jews communal lines were hardened by the emergence of the state of Israel, the hostility it evoked from most Arab states, and its aggressive efforts at ingathering. The self-consciously Islāmic government in Iran has also introduced a religious intolerance that, while it is discouraged by the Sharīʿah, is encouraged by local sentiment as well as by the staunch nationalism Iran shares with secular states. The leaders of the Islāmic Republic of Iran have associated being Iranian with being Muslim.

Farther from the centre of Islāmdom, Islām plays various roles as a minority religion. Among Turks in the Central Asian states, for example, Islām is an important source of identity. Muslims living in western Europe and the Americas are generally able to form communities and practice their religion as they will: in Canada, for example, Ismāʿīlī Muslims, under the guidance of Aga Khan IV, form a cohesive group that promotes the economic and cultural development of its members. In the United States, tenets of Islām were embraced by the founders of the American Muslim Mission (originally called Nation of Islam) in the early 1930s. As the community has developed, its leaders have increasingly emphasized the Qurʾān and Muḥammad's example as sources of authority.

Survival of Islāmic activism. Although Islāmic activism never disappeared during the years in which Muslim countries were becoming independent, other ideological orientations seemed more important between the end of World War II and the declaration of the Islāmic Republic of Iran in 1979. Many Westerners or westernized Muslims expected religion to recede as modernization progressed. Already in the 1950s, however, the Muslim Brotherhood in Egypt called for an exclusively Islāmic state in place of the secular multi-communal state that Gamal Abdel Nasser had founded. In the early 1960s new circumstances were beginning to foster increased Islāmic activity, some popular, some supported by official institutions. In these years critics of Mohammad Reza Shah Pahlavi began to rally around the exiled ayatollah Ruhollah Khomeini; the writings of ʿAli Shariʿati began to influence Muslims inside and outside Iran; and two great pan-Islāmic organizations were formed, the Muslim World League (1962) and the Organization of the Islāmic Conference (1971). Although Westerners have become most familiar with activism's violent forms, its educational, cultural, pietistic, and political dimensions have been more extensive. All these developments occurred in the wake of the formation of the Organization of Petroleum Exporting Countries in 1961 and culminated in Egypt's success in its war with Israel in 1973. The resurgence of economic and military power was not the only factor that could foster those who had maintained an interest in Islām all along. In a few parts of the Muslim world, petroleum-based prosperity promoted increased international influence and pride; elsewhere modernization was producing widespread educational and economic cleavages and populations with very low median ages. As dissatisfaction with the material failures of secular modernization grew, so did disenchantment with the Western ideologies that had undergirded it. While these other ideologies were being tried and discredited, Islām had remained relatively peripheral to public policy and thus unassailable. All the while, citizens of Muslim countries were echoing the anti-imperialist rhetoric increasing throughout the developing world.

Pan-Islāmic organizations

Situation of Muslim women. For women, modernization is especially problematic. Urged on the one hand to be liberated from Islām and thereby become modern, they are told by others to be liberated from being Western through being self-consciously Muslim. There is little information on the situation of ordinary women in premodern Islāmdom, but evidence from the modern period underscores the enormous variety of settings in which

Muslim women live and work, as well as the inability of the stereotype of meek, submissive, veiled passivity to reflect the quality of their lives. As always, Muslim women live in cities, towns, villages, and among migratory pastoral tribes; some work outside the home, some inside, some not at all; some wear concealing clothing in public, most do not; for some, movement outside the home is restricted, for most not; and, for many, public modesty is common, as it is for many Muslim men. For many, the private home and the public bath continue to be the centres of social interaction; for others, the world of employment and city life is an option. As always, few live in polygamous families. Strict adherence to the Sharīʿah's provision for women to hold their property in their own right has produced Muslim women of great wealth, in the past as well as today. Clearly, any simple description of the lives of Muslim women is misleading.

Modern Islām's unifying forces. Modern Islāmdom can appear so diverse as to defy description, yet it is also held together by stronger centripetal forces than almost any other pan-national solidarity group. The *ḥajj* attracts more than 1,000,000 Muslims annually; and, despite significant religious cleavages, Islām remains one of the least sectarian of world religions. Most Muslims live in societies in which the force of tradition is very strong and in which modernization has also penetrated to some extent. The majority of Muslims remain, as they have always been, agricultural. A very small minority are migratory pastoralists; a larger minority are village, town, and city dwellers. In all settings tradition, including religious tradition, is being drawn upon as a source of change and modernization, with the consequence that the Western equation of modernization and secularization has been severely tested and even undermined.

Yet the role of tradition varies. Some Islāmic activists rely on a kind of secularized "cultural" Islām, somewhat like cultural Judaism, that depends very little on personal piety or the observance of Islāmic law or the many customs that have come to be associated with being Muslim, while others cling to the customs associated with Islām with little awareness of Islām's more learned side. Labels such as Shīʿite, which always carried an oppositional quality, may be formerly nonessential attributes that have become salient in the wake of the success of the ayatollah Khomeini in Iran. When disadvantaged persons who happen to be Shīʿites find an opening for communal protest, or when those for whom Shīʿite theology means little find its vision of justice and radical revolution appropriate to their specific circumstances, an old label acquires a new valence.

Like any other explanatory system, Islām has always had to provide a way of talking about the world, of establishing identity in the world, and of managing the world's affairs. In performing these functions, Islām has from its inception been forced to compete with other explanatory systems for the "mental space" of its adherents and simultaneously to define its stance toward preexisting and ongoing extra-Islāmic influences. Islām continues to compete, aided unwittingly by the weaknesses of its competitors, spurred on by the freshness of its own demands for public attention, and fueled by the remarkable ability of many of its adherents to respond to the connection between the mundane and spiritual that has been the hallmark of all religious life.

BIBLIOGRAPHY

Surveys: The most visionary general work on Islāmic history is MARSHALL G.S. HODGSON, *The Venture of Islam: Conscience and History in a World Civilization,* 3 vol. (1974), which sets Islām into a world historical context. A similar but shorter work, sumptuously illustrated, is FRANCIS ROBINSON, *Atlas of the Islamic World Since 1500* (1982).

Regions of Islāmdom: PETER B. CLARKE, *West Africa and Islam: A Study of Religious Development from the 8th to the 20th Centuries* (1982); JAMIL M. ABUN-NASR, *A History of the Maghrib,* 2nd ed. (1975); CLIFFORD GEERTZ, *Islam Observed: Religious Development in Morocco and Indonesia* (1968, reissued 1971); S.M. IKRAM, *Muslim Rule in India and Pakistan, 711–1858 A.C.,* rev. ed. (1966); RAPHAEL ISRAELI, *Muslims in China: A Study in Cultural Confrontation* (1980); and NEHEMIA LEVTZION (ed.), *Conversion to Islam* (1979).

Periods and aspects of Islāmicate history: On premodern Islāmicate social structure, see ROY P. MOTTAHEDEH, *Loyalty and Leadership in an Early Islamic Society* (1980); IRA LAPIDUS, *Muslim Cities in the Later Middle Ages* (1967); and S.D. GOITEIN, *A Mediterranean Society: The Jewish Communities of the Arab World as Portrayed in the Documents of the Cairo Geniza,* 4 vol. (1967–83). HAMILTON A.R. GIBB, *Studies on the Civilization of Islam* (1962, reissued 1982), is a collection of interpretive articles on history, historiography, literature, and philology. RENÉ GROUSSET, *The Empire of the Steppes: A History of Central Asia* (1970; originally published in French, 1939); and JOHN J. SAUNDERS, *The History of the Mongol Conquests* (1971), deal with the Mongol conquests. JOHN J. SAUNDERS (ed.), *The Muslim World on the Eve of Europe's Expansion* (1966), combines primary sources on the last three great empires; and the most comprehensive account of modern Islām, with an especially fine treatment of the 18th century, is JOHN OBERT VOLL, *Islam, Continuity and Change in the Modern World* (1982). On Muslim women, see, for example, LOIS BECK and NIKKI KEDDIE (eds.), *Women in the Muslim World* (1978); ELIZABETH WARNOCK FERNEA and BASIMA QATTAN BEZIRGAN (eds.), *Middle Eastern Muslim Women Speak* (1977, reprinted 1984); and JANE I. SMITH (ed.), *Women in Contemporary Muslim Societies* (1980).

Collections of primary sources in English translation: ERIC SCHROEDER, *Muhammad's People* (1955); ARTHUR JEFFERY (ed.), *A Reader of Islam* (1962, reprinted 1980); JOHN ALDEN WILLIAMS (ed.), *Islam* (1961, reissued 1967), and *Themes of Islamic Civilization* (1971, reprinted 1982); WILLIAM H. MCNEILL and MARILYN ROBINSON WALDMAN, *The Islâmic World* (1973, reprinted 1983); JAMES KRITZECK, *Anthology of Islamic Literature* (1964, reissued 1975); and BERNARD LEWIS (ed.), *Islam: From the Prophet Muhammad to the Capture of Constantinople,* 2 vol. (1974, reissued 1976).

Major reference works: The Encyclopaedia of Islam, 5 vol. (1913–36), and a new edition, of which 5 vol. appeared from 1960 to 1986; *The Shorter Encyclopaedia of Islam* (1953, reprinted 1974), with articles culled from the *Encyclopaedia of Islam; The Cambridge History of Islam,* 2 vol. (1970, reprinted in 4 vol., 1980); JEAN SAUVAGET, *Jean Sauvaget's Introduction to the History of the Muslim East: A Bibliographical Guide* (1965, reprinted 1982; originally published in French, 2nd ed., 1961), a dated but still useful annotated bibliographic guide; and CLIFFORD EDMUND BOSWORTH, *The Islamic Dynasties: A Chronological and Genealogical Handbook,* rev. ed. (1980). JEAN JACQUES WAARDENBURG, *L'Islam dans le miroir de l'Occident,* 3rd rev. ed. (1970); and EDWARD W. SAID, *Orientalism* (1978, reissued 1979), are critiques of Western approaches to Islām. (M.R.W.)

Israel

Israel (in full: State of Israel; Hebrew: Medinat Yisra'el; Arabic: Isrā'īl) is a country of the Middle East situated at the eastern end of the Mediterranean Sea. It is bounded to the north by Lebanon, to the northeast by Syria, to the east and southeast by Jordan, to the southwest by Egypt, and to the west by the Mediterranean Sea. The total area is 7,992 square miles (20,700 square kilometres) excluding East Jerusalem and other territories occupied in the 1967 war. Jerusalem is the capital.

Israel lies at the centre of a region that is often called Palestine. It is a small country with a relatively diverse topography, consisting of a lengthy coastal plain, highlands in the north and central regions, and the Negev desert in the south. Running the length of the country from north to south along its eastern border is the northern terminus of the Great Rift Valley.

The State of Israel is the only Jewish nation in the modern period, and the region that now falls within its borders has a lengthy and rich history that dates from pre-biblical times. The desire for a Jewish homeland prompted a small number of Jews to immigrate to Palestine beginning in the late 1800s, a migration that grew dramatically during the second quarter of the 20th century with the increased persecution of Jews worldwide. This vast influx of Jewish immigrants into the region, however, caused tension with the native Palestinian Arabs. Israel fought a series of wars against neighbouring Arab states during the next 35 years, which resulted in ongoing disputes over territory and the status of refugees. Despite continuing tensions, Israel concluded peace treaties with several neighbouring Arab states during the final quarter of the 20th century.

The article is divided into the following sections:

Physical and human geography

THE LAND

Relief. Despite its small size, about 290 miles (470 kilometres) north-to-south and 85 miles east-to-west at its widest point, Israel has four geographic regions—the Mediterranean coastal plain, the hill regions of northern and central Israel, the Great Rift Valley, and the Negev—and a wide range of physical features and microclimates.

The coastal plain is a narrow strip about 115 miles long that widens to about 25 miles in the south. A sandy shoreline with many beaches borders the Mediterranean coast. Inland to the east, fertile farmland is giving way to growing agricultural settlements and the cities of Tel Aviv and Haifa and their suburbs.

Mountains of Galilee

In the north of the country, the mountains of Galilee constitute the highest part of Israel, reaching an elevation of 3,963 feet (1,208 metres) at Mount Meron (Arabic: Jebel Jarmaq). These mountains terminate to the east in an escarpment overlooking the Great Rift Valley. The mountains of Galilee are separated from the hills of the Israeli-occupied West Bank to the south by the fertile Plain of Esdraelon (Hebrew: 'Emeq Yizre'el), which, running approximately northwest to southeast, connects the coastal plain with the Great Rift Valley. The Mount Carmel range, which culminates in a peak 1,791 feet high, forms a spur reaching northwest from the highlands of the West Bank, cutting almost to the coast of Haifa.

The Great Rift Valley, a long fissure in the Earth's crust, begins beyond the northern frontier of Israel and forms a series of valleys running generally south, the length of the country, to the Gulf of Aqaba. The Jordan River, which marks part of the frontier between Israel and Jordan, flows southward through the rift from Dan on Israel's northern frontier, where it is 500 feet above sea level, first into the Ḥula Valley (Hebrew: 'Emeq Ḥula), then into the freshwater Lake Tiberias, also known as the Sea of Galilee (Hebrew: Yam Kinneret), which lies 686 feet below sea level. The Jordan continues south along the eastern edge of the West Bank—now through the Jordan Valley (Hebrew: 'Emeq HaYarden)—and finally into the highly saline Dead Sea, which, at 1,312 feet (400 metres) below sea level, is the lowest point of a natural landscape feature on the Earth's surface. South of the Dead Sea, the Jordan continues through the rift, where the rift forms the 'Arava Valley (Hebrew: "savannah"), an arid plain that extends to the Red Sea port of Elat.

The sparsely populated Negev comprises the southern half of Israel. Arrow-shaped, this flat, sandy desert region narrows toward the south, where it becomes increasingly arid and breaks into sandstone hills cut by wadis, canyons, and cliffs before finally coming to a point where the 'Arava reaches Elat.

Drainage. The principal drainage system comprises Lake Tiberias and the Jordan River. Other rivers in Israel are the Yarqon, which empties into the Mediterranean near Tel Aviv; the Qishon, which runs through the western part of the Plain of Esdraelon to drain into the Mediterranean at Haifa; and a small section of the Yarmūk, a tributary of the Jordan that flows west along the Syria-Jordan border. Most of the country's remaining streams flow seasonally as wadis. The rivers are supplemented by a spring-fed underground water table that is tapped by wells.

Soils. The coastal plain is covered mainly by alluvial soils. Parts of the arid northern Negev, where soil development would not be expected, have windblown loess soils because of proximity to the coastal plain. The soils of Galilee change from calcareous rock in the coastal plain, to Cenomanian and Turonian limestone (deposited from 99 to 89 million years ago) in Upper Galilee, and to Eocene formations (those dating from 54.8 to 33.7 million years ago) in the lower part of the region. Rock salt and gypsum are abundant in the Great Rift Valley. The southern Negev is mainly sandstone rock with veins of granite.

Climate. Israel has a wide variety of climatic conditions, caused mainly by the country's diverse topography. There are two distinct seasons: a cool, rainy winter (October–April) and a dry, hot summer (May–September). Along the coast, sea breezes have a moderating influence in summer, and the Mediterranean beaches are popular. Precipitation is light in the south, amounting to about 1 inch (25 millimetres) per year in the 'Arava Valley south of the Dead Sea, while in the north it is relatively heavy, up to 44 inches a year in the Upper Galilee region. In the large cities, along the coastal plain, annual rainfall averages about 20 inches per year. Precipitation occurs on about 60 days during the year, spread over the rainy season.

Average annual temperatures vary throughout Israel based on elevation and location, with the coastal areas adjacent to the Mediterranean Sea having milder temperatures—ranging from about 84° F (29° C) in August to about 61° F (16° C) in January—and higher rates of humidity than areas inland, especially during the winter. Likewise, higher elevations, such as Upper Galilee, have cool nights, even in summer, and occasional snows in the winter. However, the coastal city of Elat, in the south, despite its proximity to the Red Sea, is closer to the climate of the Jordan and ʿArava valleys and the Negev, which are hotter and drier than the northern coast; there, daytime temperatures reach about 70° F (21° C) in January and may rise as high as 114° F (46° C) in August, when the average high is 104° F (40° C).

Plant and animal life. Natural vegetation is highly varied, and more than 2,800 plant species have been identified. The original evergreen forests, the legendary "cedars of Lebanon," have largely disappeared after many centuries of timber cutting for shipbuilding and to clear land for cultivation and goat herding; they have been replaced by second-growth oak and smaller evergreen conifers. The hills are mostly covered by maquis, and wildflowers bloom profusely in the rainy season. Only wild desert scrub grows in the Negev and on the sand dunes of the coastal plain.

Animal life is also diverse. Mammals include wildcats, wild boars, gazelles, ibex, jackals, hyenas, hares, coneys, badgers, and tiger weasels. Notable among the reptiles are geckos and lizards of the genus *Agama* and vipers such as the carpet, or saw-scaled, viper (*Echis carinatus*). More than 380 species of birds have been identified in the region, including the partridge, tropical cuckoo, bustard, sand grouse, and desert lark. There are many kinds of fish and insects, and locusts from the desert sometimes invade settled areas. Several regions have been set aside as nature reserves, notably parts of the ʿArava in the south and Mount Carmel, Mount Meron, and the remains of the Ḥula Lake and marshes in the north. The Mediterranean coast and the Jordan and ʿArava valleys are important routes for migratory birds.

Settlement patterns. Jewish immigration in the 20th century greatly altered the settlement pattern of the country. The first modern-day Jewish settlers established themselves on the coastal plain in the 1880s. Later they also moved into the valleys of the interior and into parts of the hill districts, as well as into the Negev. Small cities such as Haifa and Jerusalem grew in size, and the port of Jaffa (Yafo) sprouted a suburb, Tel Aviv, which grew into the largest city in Israel. Jewish immigrants also settled those areas of the coastal plain, the Judaean foothills, and the Jordan and ʿArava valleys evacuated by Palestinians during the war of 1948, thereby becoming the majority in many areas previously inhabited by Arabs. Although the majority of the Bedouin of the Negev left the region when Israel incorporated the territory, the desert has continued to be largely the domain of the Arab nomads who remained or returned following the end of the fighting.

Early Jewish settlers

The non-Jewish population is concentrated mainly in Jerusalem (about one-fifth of the residents of the city) and in the north, where Arabs constitute a substantial part of the population of Galilee.

Jerusalem, perched high among the Judaean hills, is one of the great cities of the world, with a long history, unique architecture, and rich archaeological heritage. It is the capital of Israel, and its walled Old City is divided into four quarters—Muslim, Jewish, Christian, and Armenian—symbolizing its spiritual significance to the region's major religious and ethnic groups.

Rural settlement. The rural population, defined as residents of settlements with fewer than 2,000 people, amounts to less than one-tenth of the nation's total inhabitants. About one-tenth of the Jewish population is rural, of whom more than half are immigrants who arrived after 1948. The Jewish rural settlements are organized into kibbutzim (2 percent of the total population), moshavim (3 percent), and agricultural communities or individually owned farms engaged in private production. The kibbutzim and moshavim pioneered settlement in underdeveloped areas, performed security functions in border areas, and contributed substantially to the nation's ability to absorb new immigrants in the early years of the state. There has been a growing tendency among farmers to practice intensive cultivation, to diversify crops, and to shift from smallholdings to large farms.

Only a tiny fraction of the Arab population lives in rural areas. Those who do are divided between the Bedouin—many of whom now reside in permanent settlements—and residents of small agricultural villages. Thousands of Arabs from Israeli villages and the Israeli-occupied territories of the West Bank and Gaza have found employment in Israel. This ready labour pool, together with increased agricultural mechanization, has led to a drop in the number of Jewish agricultural workers. In Arab villages, fewer than half of the adult labourers, both men and women, are engaged in working the land, either as tenants or small landowners.

Urban settlement. The great majority of the population, both Jewish and Arab, reside in urban areas. As the industrial and service sectors of the economy have grown, the two large conurbations of Tel Aviv–Yafo and Haifa, along the coastal plain, now house more than half of the country's population, while the cities of Jerusalem and Beersheba contain another one-fourth. The government has made great efforts to prevent the population from becoming overconcentrated in these areas, overseeing in both the north and south the development of new towns occupied largely by the country's most recent immigrants.

New towns

The major urban centres inhabited by Arabs include cities and towns with both Arab and Jewish populations—such as Jerusalem, Haifa, ʿAkko, Lod, Ramla, and Yafo—and towns with predominantly Arab populations, including Nazareth in Galilee, where a mainly Jewish suburb is nearly equal in population to the Arab city.

THE PEOPLE

Religious and ethnic groups. Jews constitute about four-fifths of the total population of Israel. Almost all the rest are Palestinian Arabs, of whom most (roughly three-fourths) are Muslim; the remaining Arabs are Christians and Druze, who each comprise only a small fraction of the total population. Arabs constitute the overwhelming majority of the occupied territories of the West Bank and Gaza. (For information on Palestinians residing outside Israel, *see* PALESTINE).

Jews. The Jewish population is diverse. Jews from throughout the world have been immigrating to this area since the late 19th century. Differing in ethnic origin and culture, they brought with them languages and customs from a variety of countries. The Jewish community today includes survivors of the Holocaust, offspring of those survivors, and émigrés escaping anti-Semitism. The revival of Hebrew as a common language and a strong Israeli national consciousness have facilitated the assimilation of newcomers to Israel but not completely eradicated native ethnicities. For example, religious Jews immigrating to Israel generally continue to pray in synagogues established by their respective communities.

Religious Jewry in Israel constitutes a significant and articulate section of the population. As such, it is often at odds with a strong secular sector that seeks to prevent religious bodies and authorities from dominating national life. The two main religious-ethnic groupings are those Jews from central and eastern Europe and their descendants who follow the Ashkenazic traditions and those Jews from the Mediterranean region and North Africa who follow the Sephardic. There are two chief rabbis in Israel, one Ashkenazi and one Sephardi. Tension is frequent between the two groups, largely because of their cultural differences and the social and political dominance of the Ashkenazim in Israeli society. Until recently, it was generally true that the Sephardim tended to be poorer, less educated, and less represented in higher political office than the Ashkenazim.

The Karaites are a Jewish sect that emerged in the early Middle Ages. Several thousand members live in Ramla and, more recently, in Beersheba and Ashdod. Like other religious minorities, they have their own religious courts

© 2002 Encyclopædia Britannica, Inc.

MAP INDEX

Political subdivisions

Central 32 05 N 34 55 E
Haifa 32 35 N 35 00 E
Jerusalem 31 45 N 35 00 E
Northern 32 50 N 35 20 E
Southern 30 40 N 34 50 E
Tel Aviv 32 05 N 34 48 E

Cities and towns

'Afula 32 36 N 35 17 E
'Akko 32 55 N 35 05 E
'Arad 31 15 N 35 13 E
Ashdod 31 49 N 34 39 E
Ashdot Ya'aqov 32 40 N 35 35 E
Ashqelon 31 40 N 34 35 E
'Atlit 32 41 N 34 56 E
Bāqa el-Gharbiyya 32 25 N 35 03 E
Bat Yam 32 01 N 34 45 E
Beersheba (Be'er Sheva') 31 14 N 34 47 E
Bet Guvrin 31 36 N 34 54 E
Bet She'an 32 30 N 35 30 E
Bet Shemesh 31 45 N 35 00 E
Binyamina 32 31 N 34 57 E
Dan 33 14 N 35 39 E
Dimona 31 04 N 35 02 E
Elat 29 33 N 34 57 E
'En Gedi 31 27 N 35 23 E
'En H̱arod 32 33 N 35 23 E
'En Yahav 30 38 N 35 11 E
Gedera 31 49 N 34 46 E
Gesher ha-Ziw 33 02 N 35 06 E
Giv'atayim 32 04 N 34 48 E
H̱adera 32 26 N 34 55 E
Haifa (H̱efa) 32 50 N 35 00 E
H̱anita 33 05 N 35 10 E
Haẕeva 30 48 N 35 15 E
H̱azor 32 59 N 35 33 E
H̱efa, see Haifa
Herzliyya 32 10 N 34 51 E
H̱olon 32 01 N 34 46 E
H̱orvot Dor 32 37 N 34 55 E
Jerusalem (Yerushalayim) 31 46 N 35 14 E
Kafr Yāsīf 32 57 N 35 10 E
Karmi'el 32 55 N 35 18 E
Kefar Blum 33 10 N 35 36 E
Kefar H̱ittim 32 48 N 35 30 E
Kefar Sava 32 10 N 34 54 E
Lod 31 58 N 34 54 E
Mash'abbe Sade 31 00 N 34 47 E
Mazkeret Batya 31 51 N 34 50 E
Metulla 33 16 N 35 35 E
Miẕpe Ramon 30 36 N 34 48 E
Nahariyya 33 00 N 35 05 E
Nazareth (Naẕerat) 32 42 N 35 18 E
Nes Ẕiyyona 31 55 N 34 48 E
Netanya 32 20 N 34 51 E
Nir Yiẕhaq 31 14 N 34 22 E
Ofaqim 31 17 N 34 37 E
Or 'Aqiva 32 30 N 34 55 E
Pardes H̱anna 32 28 N 34 58 E
Petaẖ Tiqwa 32 05 N 34 53 E
Qiryat Ata 32 48 N 35 06 E
Qiryat Gat 31 36 N 34 46 E
Qiryat Mal'akhi 31 44 N 34 44 E
Qiryat Shemona 33 13 N 35 34 E
Qiryat Yam 32 51 N 35 04 E
Ra'ananna 32 11 N 34 53 E
Rama 32 56 N 35 22 E
Ramat Gan 32 05 N 34 49 E
Ramla 31 55 N 34 52 E
Reẖovot 31 54 N 34 49 E
Rishon le-Ẕiyyon 31 58 N 34 48 E
Rosh ha-'Ayin 32 06 N 34 57 E
Safad, see Ẕefat
Sederot 31 31 N 34 35 E
Sedot Yam 32 29 N 34 53 E
Shefar'am 32 48 N 35 10 E
Ṯaiyiba, et- 32 16 N 35 01 E
Tel Aviv–Yafo 32 04 N 34 46 E
Tiberias (Teverya) 32 47 N 35 32 E
Tira, et- 32 14 N 34 57 E
Tirat Karmel 32 46 N 34 58 E
Umm el-Fahm 32 31 N 35 09 E
Yad Mordekhay 31 35 N 34 33 E
Yavne 31 53 N 34 45 E
Yehud 32 02 N 34 53 E
Yerushalayim, see Jerusalem
Yotvata 29 53 N 35 03 E
Ẕefat (Safad) 32 58 N 35 30 E
Zikhron Ya'aqov 32 34 N 34 57 E

Physical features and points of interest

Aqaba, Gulf of 29 15 N 34 45 E
'Arava, Wadi ha- 30 58 N 35 24 E
'Arava Valley, ha- 30 10 N 35 10 E
Besor, Wadi 31 28 N 34 22 E
Bet Sa'ida Nature Reserve 32 52 N 35 38 E
Caesarea (H̱orbat Qesari), *historical site* 32 30 N 34 53 E
Carmel, Mount (Har Karmel) 32 44 N 35 02 E
Dead Sea (Yam ha-Melaẖ) 31 30 N 35 30 E
Esdraelon, Plain of ('Emeq Yizre'el) 32 36 N 35 14 E
Gadol Depression, ha- 30 56 N 34 59 E
Galilee (ha-Galil), *region* 32 54 N 35 20 E
Galilee, Sea of, see Tiberias, Lake
Gaza Strip, *region* 31 25 N 34 20 E
Gilboa', Mount 32 29 N 35 25 E
H̱adera, *river* 32 27 N 34 53 E
Hai Bar Reserve 29 50 N 34 58 E
Haifa, Bay of (Mifraẕ H̱efa) 32 53 N 35 03 E
Har Karmel, see Carmel, Mount
H̱emar, Wadi 31 08 N 35 22 E
H̱iyyon, Wadi 30 12 N 35 07 E
H̱ula Basin 33 08 N 35 37 E
Jordan (ha-Yarden), *river* 31 46 N 35 33 E
Judaea (Yehuda), *region* 31 35 N 35 00 E
Judaea, Hills of 31 40 N 35 10 E
Kinneret, Yam, see Tiberias, Lake
Kinneret-Negev Conduit 32 52 N 35 32 E
Masada (H̱orvot Meẕada), *historical site* 31 19 N 35 21 E
Mediterranean Sea 32 30 N 34 30 E
Melaẖ, Yam ha-, see Dead Sea
Meron, Mount 33 00 N 35 25 E
Meẕada, H̱orvot, see Masada
Midbar Yehuda, see Wilderness of Judaea
Mount Carmel National Park 32 42 N 35 03 E
Mount Meron Nature Reserve 32 59 N 35 26 E
Negev, *region* 30 30 N 34 55 E
Niẕẕana, Wadi 30 57 N 34 23 E
Paran, Wadi 30 24 N 35 10 E
Qesari, H̱orbat, see Caesarea
Qishon, *river* 32 49 N 35 02 E
Ramon, Mount 30 30 N 34 38 E
Ramon, Wadi 30 36 N 34 55 E
Samaria, *region* 32 15 N 35 10 E
Shivta (Subeita), *historical site* 30 53 N 34 38 E
Soreq, *river* 31 56 N 34 42 E
Subeita, see Shivta
Tabor, Mount (Har Tavor) 32 41 N 35 23 E
Tel 'Arad, *historical site* 31 17 N 35 08 E
Tel H̱ay, *historical site* 33 14 N 35 35 E

Tiberias, Lake (Yam Kinneret; Sea of Galilee) . . . 32 48 N 35 35 E
'Uvda Valley . . . . . . 29 57 N 34 57 E
Wilderness of Judaea (Midbar Yehuda), *region* . . . . 31 30 N 35 18 E
"Yahudia Forest" Nature Reserve . . . 32 56 N 35 43 E
Yarden, ha-, see Jordan
Yarmuk, ha-, *river* . . 32 38 N 35 34 E
Yarqon, *river* . . . . . . 32 06 N 34 47 E
Yehuda, see Judaea
Yizre'el, 'Emeq, see Esdraelon, Plain of
Ẕin, Wadi . . . . 30 57 N 35 19 E

and communal organizations. Considered part of Jewish society, they have maintained their separate identity by resisting intermarriage and preserving their religious rites based on the Torah as the sole source of religious law.

Other Jewish communities

Samaritans trace their roots to those Jews not dispersed when the Assyrians conquered Israel in the 8th century BCE. About half of the few hundred surviving members of the Samaritan community live near Tel Aviv in the town of Ḥolon. The rest live on Mount Gerizim (Arabic: Jabal al-Ṭūr), near Nābulus in the West Bank. They preserve their separate religious and communal organizations and speak Arabic but pray in an archaic form of Hebrew. They participate in national life as part of the Jewish population.

Arabs. Arabs constitute the largest single minority in Israel, and though most are Muslims of the Sunnite branch, Arab Christians form a significant minority, particularly in the Galilee region in northern Israel. Israel's Arabs speak a dialect of Levantine Arabic and learn Modern Standard Arabic in school. An increasing number also avail themselves of higher education within Israel's public schools and colleges, and many younger Arabs are now bilingual in Hebrew. Although most Israeli Arabs consider themselves Palestinians, all are full Israeli citizens with political and civil rights that are, with the exception of some limitations on military service, equal to those of Israeli Jews. Many Arabs participate actively in the Israeli political process, and several Arab political parties have members in the Israeli Knesset. Despite this inclusiveness, however, many Israeli Arabs still see themselves as living in an occupied state, and suspicions and antagonism persist.

The overwhelming majority of Israel's Muslims are Arabs. Like all other religious communities, Muslims enjoy considerable autonomy in dealing with matters of personal status, and they have separate religious courts for issues such as marriage, divorce, and inheritance. The state oversees their religious institutions. Israel's Bedouin, roughly one-tenth of the Arab population, are Muslim.

Most Christians in Israel are Arabs, and the Greek Catholic and Greek Orthodox churches are the largest denominations. Apart from the Greek Orthodox church, which has a patriarchate in Jerusalem, each church is dependent to a degree on a supreme hierarch abroad. These communities include Roman Catholics and Uniates (Melchites, Maronites, Chaldean Catholics, Syrian Catholics, and Armenian Catholics).

The Druze, who live in villages in Galilee and around Mount Carmel, have traditionally formed a closed, tight-knit community and practice a secretive religion founded in 11th-century Fāṭimid Egypt. Israel has recognized the Druze as a separate Arab community since 1957, and Israeli Druze serve in the armed forces.

Other groups. The Bahā'ī faith, a universal religion founded in Iran in the mid-19th century, is the only religion other than Judaism to have its world centre in Israel. A teaching centre, archive building, shrine, and administrative headquarters are located on Mount Carmel in Haifa. There are a few hundred adherents in Israel, most of whom are employed at the centre in Haifa.

The Circassians, who are Sunnite Muslims, emigrated from the Caucasus region in the 1870s. They number a few thousand and live in villages in Galilee, preserving their native language and traditions. Older Circassians speak Arabic as well as the Circassian language, but members of the younger generation speak Hebrew.

Demographic trends. The most significant demographic issue in Israel since its establishment has been Jewish immigration. In 1948 the Jewish population of Israel was about 670,000; this number increased to more than 1,000,000 the next year as a result of immigration. Between 1949 and 1997 about 2,350,000 Jewish immigrants entered the country; about 700,000 to 750,000 Jews left it, although some later returned. The total number of immigrants includes more than 320,000 Soviet Jews who came to Israel in 1989–91 and have continued to arrive at the rate of about 50,000 per year. Nearly 28,000 Ethiopian Jews immigrated in 1990–92, adding to an earlier migration of 11,000 in 1984–85. The largest proportion of Jews trace their roots to Europe (including the former Soviet Union) and North America, though some also hail from Africa (mostly North Africa), Asia, and the Middle East.

Jewish immigration

While the Jewish population has grown more from immigration than from natural increase since that time, the Arab population has grown mainly through high birth rates, which are markedly higher than among Israel's Jews. Overall, the population is youthful, with about one-fourth being 15 years old or younger. Life expectancy is among the highest in the world: some 80 years for women and 77 years for men.

THE ECONOMY

The large influx of well-trained and Western-educated European and North American immigrants contributed greatly to a rapid rise in Israel's gross national product (GNP) after 1948. Although most of them had to change occupations, a nucleus of highly skilled labour, in combination with the country's rapid founding of universities and research institutes, facilitated economic expansion. The country obtained large amounts of capital, which included gifts from world Jewry, reparations from the Federal Republic of Germany for Nazi crimes, grants-in-aid from the U.S. government, and capital brought in by immigrants. Israel has supplemented these forms of revenue with loans, commercial credits, and foreign investment.

The goals of Israel's economic policy are continued growth and the further integration of the country's economy into world markets. Israel has made progress toward these goals under difficult conditions, such as a rapid population increase, a boycott by neighbouring Arab countries (except Egypt from 1979 and Jordan from 1994), heavy expenditure on defense, a scarcity of natural resources, and a small domestic market. Despite these obstacles, Israel has achieved a high standard of living for most of its residents, the growth of substantial industrial export and tourism sectors, and world-class excellence in advanced technologies and science-based industry. However, this economic progress has not been uniform. Israeli Arabs are generally at the lower rungs of the economic ladder, and there are substantial economic divisions among Israeli Jews, mainly between the Sephardim and Ashkenazim.

Privatization of the economy

Government policy dating from the late 1970s has been directed toward privatization. The private, governmental, and, to a limited extent, cooperative sectors all coexist in an economy that supports both the broad objectives of state policy and individual enterprise.

Tax rates in Israel are among the highest in the world, with income, value-added, customs and excise, land, and luxury taxes being the main sources of revenue. The government has gradually raised the proportion of indirect taxes since the late 1950s. Taxation approaches two-fifths of the value of GNP and is about one-fourth of average household income.

The General Federation of Labour in Israel (Histadrut) is the largest labour union and voluntary organization in the country. It once was also one of the largest employers in Israel and owner or joint owner of a wide range of industries, but by the mid-1990s it had sold most of its holdings to private investors. The Manufacturers' Association of Israel and the Farmers' Union represent a large number of the country's employers.

Resources. *Mineral resources.* Mineral resources include potash, bromine, and magnesium, the last two deriving from the waters of the Dead Sea. Copper ore is located in the 'Arava, phosphates and small amounts of gypsum in the Negev, and some marble in Galilee. Israel began limited petroleum exploitation in the 1950s, and small oil deposits have been found in the northern Negev and south of Tel Aviv. The country also has reserves of natural gas in the northern Negev northeast of Beersheba and offshore in the Mediterranean.

Power. The power industry is nationalized, and electricity is generated principally from coal- and oil-burning thermal stations. The government has encouraged intensive rural electrification and has provided electricity for agriculture and industry at favourable rates.

The Israel Atomic Energy Commission was established in 1952 and has undertaken a comprehensive survey of the country's natural resources and trained scientific and technical personnel. A small atomic reactor for nuclear research was constructed with American assistance south of Tel Aviv. A second reactor, built in the Negev with French help, is used for military weapons research.

Agriculture, forestry, and fishing. Early Israeli society was strongly committed to expanding and intensifying agriculture in Palestine. As a result, a rural Jewish agrarian sector emerged that included two unique forms of farming communities, the kibbutz and the moshav. The amount of irrigated land increased dramatically and, along with extensive farm mechanization, was a major factor in raising the value of Israel's agricultural production.

These improvements have contributed to a great expansion in cultivating citrus and such industrial crops as peanuts (groundnuts), sugar beets, and cotton, as well as vegetables and flowers. Dairying has also increased considerably in importance. Israel produces the major portion of its food supply and must import the remainder.

Scarcity of water

The main problem facing agriculture is the scarcity of water. Water is diverted through pipelines from the Jordan and Yarqon rivers and from Lake Tiberias to arid areas in the south. Because almost all of the country's current water resources have been fully exploited, further agricultural development involves increasing yields from land already irrigated, obtaining more water by cloud seeding, reducing the amount of evaporation, desalinizing seawater, and expanding desert farming in the Negev by drawing on brackish water found underground. Israel has perfected drip-irrigation methods that conserve water and optimize fertilizer use.

Only a limited quantity of fish is available off Israel's Mediterranean and Red Sea coasts, and Israeli trawlers sail to the rich fishing grounds in the Indian Ocean off the Ethiopian coast and engage in deep-sea fishing in the Atlantic Ocean. Inland, fishpond production meets much of the domestic demand.

Industry. For more than 40 years local demand fueled Israeli industrial expansion, as the country's population grew rapidly and the standard of living rose. More recently, world demand for Israeli advanced technologies, software, and sophisticated equipment has stimulated industrial growth. Israel's high status in new technologies is the result of its emphasis on higher education and research and development. The government also assists industrial growth by providing low-rate loans from its development budget. The main limitations experienced by industry are the scarcity of domestic raw materials and sources of energy and the restricted size of the local market.

The Shrine of the Book, which houses the Dead Sea Scrolls, at the Israel Museum in West Jerusalem.

© Buddy Mays—TRAVEL STOCK

Leo de Wys Inc./Van Phillips

Young plants being watered by drip irrigation, which uses water brought by the Israel National Water Carrier system from northern Israel to the arid land of the Negev.

Mining and quarrying. The country's mining industry supplies local demands for fertilizers, detergents, and drugs and also produces some exports. A plant in Haifa produces potassium nitrate and phosphoric acid for both local consumption and export. Products of the oil refineries at Haifa include polyethylene and carbon black, which are used by the local tire and plastic industries. The electrochemical industry also produces food chemicals and a variety of other commodities. Oil pipelines run from the port of Elat to the Mediterranean.

Manufacturing. Industrial growth has been especially rapid since 1990 in high-technology, science-based industries such as electronics, advanced computer and communications systems, software, and weapons, and these have come to command the largest share of overall manufacturing output. Other principal products include chemicals, plastics, metals, food, and medical and industrial equipment. Israel's diamond-cutting and polishing industry, centered in Tel Aviv, is the largest in the world and is a significant source of foreign exchange. The great majority of industries are privately owned. Factories producing military supplies and equipment have expanded considerably since the 1967 war—a circumstance that stimulated the development of the electronics industry.

Services. *Finance.* Israel's central bank, the Bank of Israel, issues currency and acts as the government's sole fiscal and banking agent. Its major function is to regulate the money supply and short-term banking. The Israeli currency was devalued numerous times after 1948, and the new Israeli shekel (NIS) was introduced in September 1985 to replace the earlier Israeli shekel. The government and central bank took this measure as part of a successful economic stabilization policy that helped control a rate of inflation that had grown steadily between the 1950s and mid-1980s and had skyrocketed in the 1970s.

The Israeli shekel

The banking system shows a high degree of specialization. Commercial banks are privately owned and generally are restricted to short-term business. Medium- and long-term transactions, however, are handled by development banks jointly owned by private interests and the government. The Tel Aviv Stock Exchange was established in 1953.

Tourism. Tourism has increased significantly and become an important source of foreign exchange, although its growth at times has been affected by regional strife. Visitors are drawn to the region's numerous religious, archaeological, and historic sites—such as the Western Wall and Dome of the Rock and biblical cities such as Nazareth and

Bethlehem (the latter in the West Bank)—as well as to its geographic diversity, excellent weather for leisure activities, and links to the Jewish and Palestinian Arab diasporas. There are numerous resorts in the highlands and desert and along the coast, with most tourists coming from Europe and a growing number from North America.

Trade. Access to foreign markets has been vital for further economic expansion. Israel has free trade agreements with the European Union and the United States and is a member of the World Trade Organization. These agreements and Israel's many industrial and scientific innovations have allowed the country to trade successfully despite its lack of access to regional markets in the Middle East. A central problem, however, has been the country's large and persistent annual balance-of-trade deficit.

The growth of high technology

Imports consist mainly of raw materials (including rough diamonds), capital goods, and food. Exports more than doubled in value through the 1990s and became highly diversified, originating in all the major manufacturing sectors and in agriculture. High-technology products led the list of exports, and Israel sells fruit, vegetables, and flowers throughout Europe during the off-season.

Transportation. Israel has developed a modern, well-marked highway system, and road transport is more significant to the country's commercial and passenger services than transport by rail. Bus companies provide efficient service within and between all cities and towns, supplemented by private taxis and *sheruts*—privately owned and operated shuttles—which run on urban and interurban routes.

Shipping is a vital factor both for the economy and in communications with other countries. As a result of the closing of the land frontiers following the Arab blockade of Israel, ocean and air shipping has played a major role in the transportation of supplies. Three modern deepwater ports—Haifa and Ashdod on the Mediterranean and Elat on the Red Sea—are maintained and developed by the Israel Ports and Railways Authority and are linked to the country by a combined road and rail system. Israel's shipping access routes to both the Atlantic and Indian oceans have stimulated a continuous growth of its merchant fleet and airfreight facilities.

The international airport at Lod is the country's largest. Regular flights are maintained by several international airlines, with EL AL Israel Airlines Ltd., Israel's national carrier, accounting for the largest share of the traffic. Scheduled domestic aviation and charter aviation abroad is operated by Arkia Israeli Airlines Ltd.

ADMINISTRATION AND SOCIAL CONDITIONS

Government. *Constitutional framework.* Israel does not have a formal written constitution. Instead, its system of government is founded on a series of "basic laws" plus other legislation, executive orders, and parliamentary practice. The country is a democratic republic with a parliamentary system of government, headed by a prime minister and involving numerous political parties representing a wide range of political positions.

The Knesset

Israel's lawmaking body, the Knesset, or assembly, is a single-chamber legislature with 120 members who are elected every four years (or more frequently if a Knesset vote of nonconfidence in the government results in an early election). Members exercise important functions in standing committees. Hebrew and Arabic, the country's two official languages, are used in all proceedings.

The country's prime minister is the head of government and is entrusted with the task of forming the cabinet, which is the government's main policy-making and executive body. Israel has a strong cabinet, and its members may be—but need not be—members of the Knesset.

The president, who is the head of state, is elected by the Knesset for a five-year term, which can be renewed only once. The president has no veto powers and exercises mainly ceremonial functions but has the authority to appoint certain key national officials, including state comptroller, governor of the Bank of Israel, judges, and justices of the Supreme Court.

Local and regional government. The country is divided into six districts—Central, Jerusalem, Haifa, Northern, Southern, and Tel Aviv—and into 15 subdistricts. Local government consists of municipalities, local councils (for smaller settlements), or regional rural councils. The bylaws of the councils, as well as their budgets, are subject to approval by the Ministry of the Interior. Local government elections are held every five years.

The political process. National and local elections in Israel are by universal, direct suffrage, with secret balloting. All resident Israeli citizens are enfranchised from age 18, regardless of religion or ethnicity, and candidates for election must be at least 21 years old. For national races, the system of election is by proportional representation, and each party receives the number of Knesset seats that is proportional to the number of votes it receives.

Israel's party system has traditionally been complex and volatile: splinter groups are commonly formed, and party alliances often change. Cabinets are therefore invariably coalitions, often of broad political composition, since no single party has ever been able to obtain an absolute majority in the Knesset. Electoral reform in 1992 brought about two significant changes: direct election of the prime minister and primary elections to choose lists of party candidates. The primary system enhanced participatory democracy within the parties, while the prime ministerial ballot increased the power of smaller parties, further splintering the composition of the Knesset and making governing coalitions more difficult to maintain. As a result, direct elections for the premiership were repealed in 2001, and Israel returned to its earlier practice, in which the governing coalition's leader sits as prime minister.

Failure of direct elections

Political parties are both secular and religious, with the Jewish secular parties being Zionist and ranging in orientation from left-wing socialist to capitalist, and the religious parties tending to have ethnic appeal (Sephardi or Ashkenazi). There are also several Arab parties.

Israeli-occupied Arab territories. After the 1967 war, Arab territories occupied by Israeli forces were placed under military administration. These included the territory on the west bank of the Jordan River (the West Bank) that had been annexed by Jordan in 1950, the Gaza Strip, the Sinai Peninsula region of Egypt, and the Golan Heights region of Syria. In addition, East Jerusalem (also formerly part of Jordan) was occupied by Israeli forces, and Israel took over administration of the city as a single municipality; in 1967 Israel incorporated East Jerusalem and adjoining villages and later formally annexed them—actions that have continued to be disputed abroad and hotly contested by Palestinians and neighbouring Arab nations. In 1978 the Israeli military occupied a strip of Lebanese territory adjoining Israel's northern border, from which it withdrew in 2000. Israel passed legislation effectively annexing the Golan Heights in April 1981, but completed a withdrawal from the Sinai Peninsula in April 1982 after negotiating a peace treaty with Egypt.

In May 1994, Israel began turning over control of much of the Gaza Strip and parts of the West Bank—including jurisdiction over most of the people in those areas—to the Palestinians as part of a series of agreements (generally referred to as the Oslo Accords) designed to settle outstanding grievances between the two sides. Subsequent agreements with the Palestinians met with little success, and the entire peace process broke down amid increasing violence in September 2000.

Justice. Municipal, religious, and military courts exercise a jurisdiction almost identical to that exercised by such courts during the period of the Palestine Mandate. Regional labour courts were established in 1969, and matters of marriage and divorce are dealt with by the religious courts of the various recognized communities. Capital punishment has been maintained only for genocide and crimes committed during the Nazi period.

The president appoints judges of the magistrates', district, and supreme courts, and judges hold office until mandatory retirement. The Israeli judiciary is highly independent from political influence. The country has convened special investigative panels on unusual occasions—for example, in the aftermath of the war of 1973 and following the massacre of Palestinians by Christian militiamen in Israeli-controlled sectors of Lebanon in 1982—to issue reports and allocate responsibility among political and military leaders.

Security services

The police in Israel are a branch of the Ministry of Public Security and report to a national headquarters commanded by an inspector general. The same ministry administers the nation's prison system, which is linked to a system to rehabilitate prisoners following their release. The Border Guard is a military arm of the national police and is responsible for maintaining internal security and combating terrorism. A Civil Guard, formed in 1974 by the government to prevent terrorism, consists of volunteers performing neighbourhood-watch and patrol duties.

The armed forces. The Israel Defense Forces (IDF) is generally regarded by military experts as one of the finest armed forces in the world. IDF doctrine has been shaped since Israel's founding by the country's need to stave off attack from the numerically superior and geographically advantaged forces of its hostile Arab neighbours. This doctrine encompasses the IDF's belief that Israel cannot afford to lose a single war, a goal that it feels can be attained only through a defensive strategy that includes a peerless intelligence community and early warning systems and a well-trained, rapidly mobilized reserve component combined with a strategic capability that consists of a small, highly trained, active-duty force that is able to take the war to the enemy, quickly attain military objectives, and rapidly reduce hostile forces.

An integrated organization encompassing sea, air, and land forces, the IDF consists of a small corps of career officers, active-duty conscripts, and reservists. Since the IDF depends on reservists to meet manpower requirements, it continues to be mainly a popular militia rather than a professional army. Consequently, civilian-military relations are based firmly on the subordination of the army to civilian control. Military service is compulsory for Jews and Druze, both men and women, and for Circassian men. Muslim and Christian Arabs may volunteer, although, because of security concerns, the air force and intelligence corps are closed to minorities.

Education. Schooling is obligatory and free for children between the ages of five and 15 and free, but not compulsory, for those 16 and 17. Young people between the ages of 14 and 18, however, who have not completed secondary schooling are obliged to attend special classes. Parents may choose to send children to state secular schools, state religious schools, or private religious schools. For Arab students, there is a system of schools in which Arabic is the primary language of instruction.

Institutions of higher learning

In addition to the Hebrew University of Jerusalem (1925), the Technion-Israel Institute of Technology in Haifa (1924), and the Weizmann Institute of Science in Reḥovot (1934), several institutions of higher learning have been founded since 1948, including the universities of Tel Aviv and Haifa, Bar-Ilan University (religious, located near Tel Aviv), and Ben-Gurion University of the Negev in Beersheba. The Open University of Israel (formerly Everyman's University) in Tel Aviv opened in 1974, and teachers' training colleges include two for Arabs. The language of instruction at Israeli universities is Hebrew, while the teaching system represents a mixture of European and American methods. Academic freedom in the universities is protected by Israeli law.

Health and welfare. The Ministry of Health maintains its own public and preventive health services, including hospitals and clinics, and it supervises the institutions of nongovernmental organizations. A national health insurance program assures hospitalization coverage and basic medical care for all. Several health maintenance organizations are open to all Israelis, the largest of which, Kupat Holim—with its own physicians, clinics, and hospital—is run by the Histadrut labour union and is recognized worldwide as an exemplary health care organization. Israel ranks among the most successful countries in the world in terms of the proportion of its GNP spent on health care and its rates of life expectancy and infant mortality.

The Ministry of Labour and Social Affairs supervises the service bureaus that deal with family, youth, and community welfare, as well as with rehabilitation of the handicapped. Most of these bureaus operate within local or regional government. Membership in the country's social-insurance plan is compulsory. The program provides welfare, child care and family allowances, income maintenance, disability insurance, old-age pensions, and long-term care for the elderly.

CULTURAL LIFE

The cultural milieu. There has been little cultural interchange between the Jewish and Arab sections of Israel's population, although Jews arriving in Israel from communities throughout the world, including the Arab-Muslim Middle East, have brought with them both their own cultural inheritance and elements absorbed from the majority cultures in which they dwelt over the centuries. The intermingling of the Ashkenazic, Sephardic, and Middle Eastern traditions has been of profound importance in forging modern Israel; however, the arrival of immigrants from Russia and other former Soviet republics has slowed the trend, common among immigrants from central Europe and America, toward creating a cultural synthesis embracing East, West, and native Israeli society. The revival of the Hebrew language as a vernacular in the 20th century was at the forefront of Israel's cultural development. This diverse cultural heritage but shared language, along with a common Jewish tradition, both religious and historical, form the foundation of Israel's complex culture.

The arts. The Israel Philharmonic Orchestra has earned a worldwide reputation for classical music, and Israeli artists such as violinists Itzhak Perlman and Pinchas Zukerman and pianist and conductor Daniel Barenboim have had prominent international careers. Folk dancing and popular singing enjoy widespread interest and combine foreign elements with original creative manifestations. The Sephardic, Ashkenazic, and Arab Palestinian communities have all preserved parts of their ethnic music and dance traditions. In 2000 the Education Ministry began including Israeli-Arab writers in the literature curriculum of state secular schools. Painting and sculpture are still largely influenced by European schools, but local styles have begun to emerge, and several "primitive" artists whose works depict biblical and local themes have become popular. In literature, poetry, and drama, a concentration on themes of the Diaspora is giving way to an interest in national themes, including the Holocaust. Among Israel's most distinguished writers is Shmuel Yosef Agnon (1888–1970), who received the Nobel Prize for Literature in 1966.

Israeli popular culture

Thanks to an advanced and pervasive communication infrastructure, including cable, satellite, and Internet access, Israeli popular culture is well informed and tuned to the latest international trends and performers. New Israeli pop singers and groups performing in Hebrew emerge frequently. The sound is global and is influenced by folk, rock, and all the latest styles, but the lyrics are uniquely Israeli, reflecting the concerns of the nation's youth. At the same time locally produced talk shows in Hebrew are prime-time favourites. In addition to cable and satellite access, Arab neighbourhoods and towns bristle with television antennas permitting reception from neighbouring Arab countries and making Arabic pop music widely available.

Sports. A wide variety of sports are pursued in Israel, from organized team sports such as association football (soccer) and basketball—two perennial favourites—to popular outdoor pastimes such as mountain biking, windsurfing, and scuba diving. The country hosts the quadrennial Maccabiah Games, a major cultural event for Jewish athletes worldwide.

Jewish pioneers formed the Palestine Olympic Committee in 1933, but the first Israeli team did not participate in the games until the 1952 Summer Games at Helsinki, Finland. However, Israel's Olympic team is perhaps best remembered for the tragic kidnapping and murder of 11 of its members by Palestinian terrorists at the 1972 Summer Games in Munich, West Germany.

Cultural institutions. Israel has a rich and varied range of cultural institutions, including major libraries, an art institute and artists' colonies, art museums, institutes for archaeology and folk life, theatres, concert halls and performing arts centres, and movie houses. A thriving film industry has emerged. In 1953 the Israeli government established the Academy of the Hebrew Language as the supreme authority on all questions related to the language and its usages, and it founded the Israel Academy of Sciences and Humanities in 1959. The Jewish National and University Library in Jerusalem is preeminent among the nation's several hundred libraries. Habima, Israel's national theatre, was founded in Moscow in 1917 and moved to Palestine in 1931. There are a number of other theatres in the country, some of them in the kibbutzim. Foremost among the many art galleries and museums is the Israel Museum in Jerusalem, which also houses part of the archaeological collection of the government's Department of Antiquities. The discovery of the Dead Sea Scrolls in 1947 was a powerful stimulus to biblical and historical research in the country.

Dead Sea Scrolls

Press and broadcasting. Tel Aviv is the centre of newspaper publishing in Israel. In the past newspapers were often associated with a political party, but most have now passed to private ownership. Most newspapers are written in Hebrew, but a considerable number are also published in other languages. There are hundreds of other periodicals, of which more than half are in Hebrew.

The Israel Broadcasting Authority controls and licenses the broadcasting industry. Commercial radio and broadcasting has been allowed since 1986. There are two public radio networks—one providing classical programming and the other more popular music—and an armed forces station; in addition several private radio stations have been established since 1986. Programs are broadcast mainly in Hebrew, Arabic, and English but also in a wide variety of other languages, including Yiddish, Russian, Ladino (a Spanish dialect of the Sephardic Jews), and Moghrabi (Moroccan Judeo-Arabic).

Television programming, introduced in 1966, is in Hebrew and Arabic. There are two television networks, one of which is government-owned and the other privately funded, and an educational television service. Cable and direct-dish television are available in much of the country, providing a wide range of international programming via satellite, and Internet access is widely available.

(E.E./W.L.O./Ru.A.S.)

History

This discussion focuses primarily on the modern state of Israel. For treatment of earlier history and of the country in its regional context, *see* PALESTINE.

The nation of Israel is the world's first Jewish state in two millennia. It represents for Jews the restoration of their homeland after the centuries-long Diaspora that followed the demise of the Herodian kingdom in the 1st century CE. As such, it remains the focus of widespread Jewish immigration, and more than one-third of world Jewry now lives there.

A violent birth

The country, barely half a century old, was born in the midst of war. It took Israel three decades and numerous conflicts, large and small, to achieve its first peace treaty with a neighbouring Arab country, Egypt. That process has been complicated by Israel's relations with the Palestinians, many of whom were displaced by the Arab-Israeli war in 1948 or came under Israeli rule following the Six-Day War in 1967. Only since the signing of the Oslo Accords in 1993 between Israel and the Palestine Liberation Organization (PLO) has an inclusive peace settlement with other Arab states and with the Palestinians been a possibility.

Israel's national security policy has been dominated by the prime minister and shaped by coalition politics—often disrupted by social and religious issues—that have characterized the state since its creation. While the country's two major parties, the left-wing Labour and right-wing Likud, often found a consensus on security issues, especially during crises, they retained an important difference that dates to the early years of Zionism. Both parties affirm Jewish rights to the biblical land of Israel (an area, all told, only slightly larger than the U.S. state of Vermont), but Labour has been ready, as the price of peace, to cede sovereignty to the Arabs of part of the area Israel occupied after the 1967 war, while Likud has insisted that control of that territory is vital to Israel's security. Neither political party, however, has been prepared to accept the return to Israel of large numbers of Palestinian refugees. Despite extensive peace talks following Oslo and a peace treaty with Jordan in 1994, Israel has not been able to come to an amicable peace with Syria or Lebanon or with the Palestinians. Persistent bouts of violence have dimmed hopes for peace.

Domestically, Israel moved steadily from an economy directed by the state to one that was more market-oriented. A novel feature of the earlier economy was the kibbutz, a collective settlement movement that exemplified the Labour-Zionist movement's ideals of sacrifice and leadership. After Labour lost political power to the nonsocialist Likud opposition in 1977, the kibbutz ideal began to wane and with it the socialist and secular beliefs so strong at Israel's birth. The huge influx of Sephardic Jews in the 1950s and Russian Jews in the 1980s and '90s forced more political, social, and economic changes, as these groups acquired increasing power and influence. A crucial unresolved question remained: the relationship between religion and state.

Religion and state

ORIGINS OF A MODERN JEWISH STATE

Zionism. Modern Israel springs from both religious and political sources. The biblical promise of a land for the Jews and a return to the Temple in Jerusalem were enshrined in Judaism and sustained Jewish identity through an exile of 19 centuries following the failed revolts in Judaea against the Romans early in the Common Era. By the 1800s, fewer than 25,000 Jews still lived in their ancient homeland, and these were largely concentrated in Jerusalem, then a provincial backwater of the Ottoman Empire.

In the 1880s, however, a rise in European anti-Semitism and revived Jewish national pride combined to inspire a new wave of emigration to Palestine in the form of agricultural colonies financed by the Rothschilds and other wealthy families. Political Zionism came a decade later, when the Austrian journalist Theodor Herzl began advocating a Jewish state as the political solution for both anti-Semitism (he had covered the sensational Dreyfus affair in France) and a Jewish secular identity. Herzl's brief and dramatic bid for international support from the major powers at the First Zionist Congress (August 1897) failed, but, after his death in 1904, the surviving Zionist organization under the leadership of Chaim Weizmann undertook a major effort to increase the Jewish population in Palestine while continuing to search for political assistance.

These efforts could only be on a small scale while the Ottoman Turks ruled what the Europeans called Palestine (from Palaestina, "Land of the Philistines," the Latin name given Judaea by the Romans). But in 1917, during World War I, the Zionists persuaded the British government to issue the Balfour Declaration, a document that committed Britain to facilitate the establishment of a "Jewish homeland" in Palestine. Amid considerable controversy over conflicting wartime promises to the Arabs and French, Britain succeeded in gaining the endorsement of the declaration by the new League of Nations, which placed Palestine under British mandate.

The Balfour Declaration

Immigration and conflict. The Zionist goal of Jewish statehood was violently opposed by the local Arab leaders, who saw the Ottoman defeat as an opportunity either to create their own state or to join a larger Arab entity—thus reviving the old Arab empire of early Islāmic times. British efforts to bring the Zionists and the Arabs together in a cooperative government failed, and serious disorders, escalating into organized violence, were to mark the mandate, culminating in the Arab Revolt of 1936–39. This period also marked the birth of local Jewish defense forces. The largest and most widely representative of the various militias, the Haganah ("Defense") was a branch of the Jewish Agency, the organization most responsible for bringing Jews to Israel.

The most effective of the main, pre-state militias were associated with political factions from both the right and left wings of Zionist politics. The Irgun Zvai Leumi and its even more violent splinter group, Lehi (also known as the Stern Gang), were affiliated with the ultraconservative Revisionist Party, founded by Vladimir Zev Jabotinsky. (The Revisionists withdrew from the main Zionist institutions in 1935 in protest against Jewish cooperation with the British mandate.) Another group, the Palmach, though technically an elite arm of the Haganah, was heavily influenced by a Marxist-socialist party, Achdut HaAvoda, and recruited many of its members from socialist-oriented kibbutzim. Members of these militias were to play an important role in Israeli politics for the next half century: Yigal Allon, Moshe Dayan, and Yitzhak Rabin were high-ranking members of the Haganah-Palmach, Menachem Begin led the Irgun, and Yitzhak Shamir was a prominent member of the Lehi. Three of these, Rabin, Shamir, and Begin, would later become prime ministers of Israel.

Role of early militias

Britain encouraged Jewish immigration in the 1920s, but the onset of the worldwide Great Depression of the 1930s and the flight of refugees from Nazi Germany led to a change in policy. The British government proposed the partition of Palestine into mutually dependent Arab and Jewish states. When this was rejected by the Arabs, London decided in 1939 to restrict Jewish immigration severely in the hope that it would retain Arab support against Germany and Italy. Palestine was thus largely closed off to Jews fleeing Nazi-dominated Europe during World War II. Despite this fact, the majority of the Jewish population supported the Allies during the war while seeking, when possible, clandestine Jewish immigration to Palestine. The Jewish community, which was less than 100,000 in 1919, numbered some 600,000 by the end of the war. The Arabs of Palestine had also increased under the mandate (through high birth rates and immigration) from about 440,000 to roughly 1,000,000 in 1940.

The pre-Holocaust Zionist struggle to secure international support, overcome Arab opposition, and promote immigration resumed with special fervour after 1945, when the true extent of Jewish losses in Europe became evident. In Britain, the newly elected government of Prime Minister Clement Attlee, alarmed by growing violence in Palestine between Arabs and Jewish immigrants, decided to end the mandate, but it was unable to do so in a peaceful way. Atlee and his foreign secretary, Ernest Bevin, came under pressure by the Zionists and their sympathizers, especially President Harry S. Truman in the United States, to admit the desperate remnant of European Jewry into Palestine; they were equally pressured by local and regional Arab opponents of a Jewish state to put an end to further immigration. Both sides, Arab and Jewish, violently assailed the reinforced garrison in Palestine of the war-weakened British. Finally, London turned the problem over to the newly formed United Nations (UN), and on November 29, 1947, the UN General Assembly voted to divide British-ruled Palestine into two states, one Jewish and the other Arab. This decision was immediately opposed by the Arabs who, under the ostensible leadership of Hajj Amīn al-Ḥusaynī, the grand mufti of Jerusalem, attacked Jews throughout Palestine as the British withdrew.

ESTABLISHMENT OF ISRAEL

The war of 1948. The Zionist militias gained the upper hand over the Palestinians through skill and pluck, aided considerably by intra-Arab rivalries. Israel's declaration of independence on May 14, 1948, was quickly recognized by the United States, the Soviet Union, and many other governments, fulfilling the Zionist dream of an internationally approved Jewish state. Neither the UN nor the world leaders, however, could spare Israel from immediate invasion by the armies of five Arab states—Egypt, Iraq, Lebanon, Syria, and Transjordan (now Jordan)—and the state's survival appeared to be at stake.

The Israeli forces, desperately short of arms and training, still had the advantage of having just beaten al-Ḥusaynī's irregulars, and their morale was high. David Ben-Gurion, the new prime minister, had also, soon after independence, unified the military command, although this process was bloody. When an Irgun ship called the *Altalena* attempted to land near Tel Aviv in June 1948 under conditions unacceptable to Ben-Gurion, he ordered it stopped. Troops commanded by Yitzhak Rabin fired on the vessel, killing 82 people (Menachem Begin was one of the survivors). The Irgun and Palmach finally consented to the unified command, but relations between the Labour movement Ben-Gurion had established and its right-wing opposition, founded in Jabotinsky's Revisionist Party, were poisoned for years.

The *Altalena* incident

The Arab invaders far outnumbered the Zionists but fielded only a few well-trained units. In addition, some Arab logistical lines were long, making resupply and communication difficult. The most formidable Arab force was Transjordan's British-led Arab Legion, but the Jordanian ruler, King ʿAbdullāh, had secret relations with the Zionists and strongly opposed a Palestinian state led by his enemy al-Ḥusaynī. Other states, such as Egypt and Iraq, also had different objectives, and this internal strife, disorganization, and military ineptitude prevented the Arabs from mounting a coordinated attack.

In June all sides accepted a UN cease-fire, and the nearly exhausted Israelis reequipped themselves. Fierce fighting resumed in early July and continued for months interspersed with brief truces. The Israelis drove back the Egyptian and Iraqi forces that menaced the south and central parts of the coastal plain. However, the old walled city of Jerusalem, containing the Western Wall, the last remnant of the ancient Temple destroyed by the Romans and held holy by Jews, was occupied by the Jordanians, and Jerusalem's lifeline to the coast was jeopardized. The Egyptians held Gaza, and the Syrians entrenched themselves in the Golan Heights overlooking Galilee. The 1948 war was Israel's costliest: more than 6,000 were killed and 30,000 wounded out of a population of only 780,000.

Armistice and refugees. Initial UN mediation conducted by Swedish diplomat Count Folke Bernadotte produced a peace plan rejected by all sides, and Bernadotte himself was murdered by Lehi extremists in September 1948. When Israel secured the final armistice of the war in July 1949, the new state controlled one-fifth more territory than the original partition plan had specified and rejected a return to the original partition line. Jordan occupied the West Bank, which was much of the area assigned by the UN to the stillborn Palestinian state, and more than 600,000 Arab refugees fled their homes in an exodus that had begun even before May 1948. Some were forced out by Israeli troops, notably from the towns of Lod and Ramla in the strategic area near Tel Aviv airport. The Israeli government refused to permit these refugees, who gathered under UN care in camps in Gaza, the West Bank, southern Lebanon, and Syria, to return to their homes inside Israel, and many Palestinians were to stay in these camps indefinitely.

Israel's victory in the war did not bring peace. The Arabs, who were humiliated by defeat and still bitterly divided, refused to recognize the Jewish state. In early 1949, the Arab nations announced a state of war with Israel and organized an economic and political boycott of the country.

THE BEN-GURION ERA

Emergence of a nation. The new Israeli state thus had to deal with challenges similar to those faced by the pre-1948 Zionist movement and needed foreign assistance, an effective strategy to hold off the Arabs, and massive Jewish immigration to settle the land in order to survive. All of this had to be done at once, and none of it could be possible without Israeli national unity.

Israel's first regular election in 1949 returned Ben-Gurion to power but did not give his Mapai (Labour) Party a majority. This set a pattern, and every Israeli government since independence has been formed as a coalition. Ben-Gurion sought a centrist position, condemning those to his left as pro-Soviet and those to his right as antidemocratic. He buttressed these arrangements by adding the Zionist religious parties to his largely secular coalition in what became known as the "status quo." The Orthodox Jewish religious parties backed Ben-Gurion on security issues, while Ben-Gurion supported an Orthodox monopoly over the control of marriage, divorce, conversion, and other personal status issues. Part of the status quo, however, included rejecting the idea of drafting a written constitution or bill of rights, and the Jewish content of the Jewish state thus would be defined by the rough-and-tumble of Israeli politics and the evolution of Israeli society.

The "status quo"

During the early years, Israel had to absorb a major influx of immigrants, including several hundred thousand nearly destitute Holocaust survivors and a large influx of Sephardic Jews from Arab states, who felt increasingly insecure in their home countries following the Arab defeat in 1948. As a result, the Knesset passed the Law of Return in 1950, granting Jews immediate citizenship. This law, however, proved to be controversial in later years when the question of "who is a Jew?" raised other issues in the Jewish state, including those of the immigration of non-Jewish relatives, religious conversion, and, in light of the Orthodox monopoly over such matters, the issue of who is truly qualified to be a rabbi.

No less serious was the question of ethnicity. The Sephardim, or Oriental Jews, were mostly from urban and traditional societies, and after arriving in Israel they encountered an Ashkenazic, or European, Zionist establishment intent on creating a new Israeli culture and settling these predominantly urban newcomers in rural and isolated villages and development towns. The Sephardim soon grew to resent what they regarded as a patronizing Ashkenazic elite, and eventually this was to hurt Labour at the ballot box.

Israel was impoverished, and its economy emerged from severe austerity only after 1952 when the country began to obtain substantial international aid, including grants from Jewish charities, revenue from the sale of bonds, and U.S. government assistance. Beginning in 1953, Ben-Gurion secured economic aid from what was then West Germany, a highly controversial act that was seen by many as reparations for the Holocaust.

Continuing tensions. Despite its victory in the 1948 war, Israel soon faced new and severe threats. Arab refugees infiltrated the armistice lines seeking to reclaim fields and houses. Soon, irregular Arab forces, drawn from refugee camps outside Israel's borders, began to attack Israeli villages, farms, and road traffic. Israel also contained a sizable minority of Arabs (then roughly one-sixth of the population), who were kept under military rule in certain areas until 1966 and, in some cases, were relocated away from border zones.

Israel's borders were difficult to defend; only 12 miles (19 kilometres) separated Jordanian army positions from the Mediterranean, and the road connecting Jerusalem with the rest of the Israel was within rifle range of Arab sharpshooters. Israel's potential allies, including the United States, were preoccupied with the Cold War and were willing to placate Arab leaders in order to limit Soviet influence among the Arab states, especially Egypt, which looked to Moscow for help against Britain and France, the remaining colonial powers in the region.

Relations with Jordan

Israel's best chance for peace was King ʿAbdullāh of Jordan, but in 1950 Palestinian and Arab opposition forced him to abandon a secretly negotiated nonbelligerency agreement. In July 1951, the Jordanian king, who had annexed the West Bank a year earlier, was assassinated on the Temple Mount in Jerusalem by a Palestinian. His grandson, King Ḥussein, continued ʿAbdullāh's policy of clandestine contact with Israel but never felt politically strong enough to make a separate peace.

In the period 1949–53 Arab attacks killed hundreds of Israelis, four-fifths of whom were civilian. In early 1953 Israel decided to take the offensive against Arab guerrillas who were infiltrating from Jordan and the Egyptian-run Gaza Strip. The Israel Defense Forces (IDF) escalated retaliations, fighting pitched battles not only with guerrillas but with regular Jordanian and Egyptian army units.

The Suez War. The Israeli raids humiliated Egypt's nationalist government, headed by Gamal Abdel Nasser. After he failed to obtain American arms to repel the Israeli attacks, Nasser trumped both Israel and his Western adversaries, the British and French, when in October 1955 he signed a security agreement with the Soviet Union and a major arms deal with Czechoslovakia that threatened overnight to erase Israel's tenuous margin of military superiority, especially in aircraft. He also announced a blockade of the Strait of Tiran, the outlet of Israel's southern port city of Elat.

Ben-Gurion, exhausted by political struggles, had left the premiership in late 1953 to Moshe Sharett, who hoped that vigorous international diplomacy might relieve Israel's insecurity. It did not. Ben-Gurion had a different approach, and, returning as prime minister in late 1955 after the Czech-Egyptian arms deal, he soon began to plan a preemptive attack against Egypt before that country's new weaponry gave it strategic superiority. The preparations for an Israeli attack coincided with the Anglo-French decision to regain the Suez Canal, which Nasser had nationalized in July 1956. The French brokered a secret alliance with Israel and Britain, and in October IDF troops, under the leadership of Moshe Dayan, swiftly broke the Egyptian lines in the Sinai. The Israeli attack provided the cover for a ruse in which the British and French invaded the canal zone under the pretext of protecting it. This duplicity infuriated U.S. President Dwight D. Eisenhower, who compelled the British and French governments to withdraw their troops. Israel was also compelled to return to the old armistice lines, but not before the United States had agreed to placing a UN peacekeeping force in the Sinai. U.S. Secretary of State John Foster Dulles also promised in writing that the United States would treat the Strait of Tiran as an international waterway and keep it open.

These arrangements did not lead to peace negotiations, but they did impose a calm over Israel's southern border for nearly a decade. A regional arms race began in the absence of any movement toward peace, and Shimon Peres, Ben-Gurion's deputy defense minister, found France to be a willing supplier. The French-designed nuclear reactor in Dimona was widely suspected of being the kernel of an Israeli nuclear weapons program. Israel, under the leadership of IDF Chief of Staff Yitzhak Rabin, turned its military into a highly professional organization.

LABOUR RULE AFTER BEN-GURION

Ben-Gurion stepped down as prime minister in June 1963. His efforts at building the Israeli state had also brought him into conflict with his own party's ideology, the Orthodox religious establishment, and the international Zionist movement. Gathering about him a group of younger leaders in 1965, notably Shimon Peres and Moshe Dayan, Ben-Gurion organized a new political party, Rafi, though he eventually retired from politics permanently in 1970 when that party failed to generate support.

The Six-Day War. Ben-Gurion's successor, Levi Eshkol, had much less experience in defense issues and relied heavily on Rabin. Neither the Jordanian nor the Syrian borders were quiet during the years leading up to the Six-Day War, but all Israelis were taken by surprise when in May 1967 increasingly violent clashes with Palestinian guerrillas and Syrian army forces along Lake Tiberias led to a general crisis. The Soviet Union alleged that Israel was mobilizing to attack Syria, and the Syrian government, in turn, chided President Nasser of Egypt for inaction. Nasser then mobilized his own forces, which he promptly sent into the Sinai after he ordered that UN forces there be withdrawn, and announced a blockade of the Strait of Tiran. The encirclement of Israel was complete when King Ḥussein of Jordan, despite secret Israeli pleas, felt compelled to join the Arab war coalition.

In reaction, Eshkol mobilized the IDF and sent his foreign minister, Abba Eban, on a futile trip to seek French, British, and American aid. After Rabin suffered a breakdown from exhaustion, the coalition parties forced Eshkol to appoint Moshe Dayan as defense minister and to create a national unity government that included Menachem Begin, the main opposition leader. The next day, June 5, Israeli planes destroyed the Egyptian air force on the ground in a preemptive strike that began the total rout of all Egyptian, Jordanian, and Syrian forces. Six days later, Israeli troops stood victorious along the Suez Canal, having overrun the Sinai Peninsula; on the banks of the Jordan River, after occupying the entire West Bank; and atop the Golan Heights, after driving the Syrians from that strategic position. Most significant to all involved, Israel had captured the remaining sections of Jerusalem not already under its control, including the Old City and the Western Wall.

Occupation of Arab territory

Troubled victory. Israel's triumph in the Six-Day War (the name by which this conflict became known) brought the entire biblical land of Israel under Jewish control, but the war also brought new complications: rule over more than one million additional Palestinians in the occupied territories; extended military lines along the Suez Canal, the Golan Heights, and the Jordan River that severely strained its small standing army; and strong international opposition to the expansion of Israeli control, especially the absorption of the Old City of Jerusalem.

It was not clear how military victory could be turned into peace. Shortly after the war's end Israel began that quest, but it would take more than a decade and involve yet another war before yielding any results. Eshkol's secret offer to trade much of the newly won territory for peace agreements with Egypt, Jordan, and Syria was rejected by the Arabs, who refused to negotiate directly with Israel. The UN Security Council responded by passing Resolution 242 in November, demanding that Israel withdraw from "occupied territories" and that all parties in the dispute recognize the right of residents of each state to live within "secure and recognized borders." The wording of this statement became crucial to peace negotiations for years to come. By not stating "all the occupied territories" in the English version—the only one accepted by Israel—the resolution left room for the Israelis to negotiate. The Palestinians, the residents of these territories, were mentioned only as refugees, it being presumed that Jordan would represent them.

Palestinian relations under Labour

Nearly two years of fruitless mediation ensued while Israel held the occupied territories with a minimum of force, relied on its air power to deter Arab attack, and—adhering to Dayan's light-handed occupation policy—disturbed the Palestinian population under military rule as little as possible. The Israelis left the Temple Mount in Jerusalem, the local Arab institutions, and indeed the Jordanian legal code throughout the West Bank in the hands of the Palestinians, just as they left Egyptian regulations in place in Gaza.

The Israeli and Palestinian economies were to develop strong links over the next decades, as the underemployed Arab workforce in the occupied territories gravitated to Israeli industries that were chronically short of unskilled labour. Eventually more than 150,000 workers would make the daily commute to Israel, returning to the West Bank and Gaza at night. While the export of Israeli goods to the occupied territories became lucrative, it formed but a small part of the exchange between the two sides.

After the Arabs rejected a quick peace, Yigal Allon, a leading Labour politician and a hero of the 1948 war, devised a plan to settle Jews in strategic areas of the West Bank, Gaza, Sinai, and the Golan Heights. Israel also enlarged the municipal boundaries of Jerusalem and developed new neighbourhoods in order to establish a Jewish majority in the capital; and in the Old City, the government reconstructed the historic Jewish quarter. However, except for East Jerusalem, where the Jewish population increased dramatically in the years of Labour dominance, by 1977 only about 5,000 Israelis lived in these so-called strategic settlements. Other Israelis, guided by the prominent Rabbi Zvi Yehuda Kook, believed that settlement everywhere in the biblical land of Israel would hasten the messianic era. Israelis of this mind established the Gush Emunim ("Bloc of Believers") organization in the West Bank city of Hebron in 1968.

The Gush Emunim

The war of attrition. In early 1969 Egypt began what became known as the "war of attrition" against Israel. Using heavy artillery, new MiG aircraft, Soviet advisers, and an advanced Soviet-designed surface-to-air missile system, the Egyptians inflicted heavy losses on the Israelis. Golda Meir, who became Israel's prime minister following Eshkol's sudden death in February 1969, escalated the war by ordering massive air raids deep into Egypt. Israel was also beset by guerrilla raids from Jordan, launched by Yāsir ʿArafāt's Palestine Liberation Organization (PLO). These attacks were often on nonmilitary targets, and Israel soon stamped the PLO as a terrorist organization and refused to negotiate with it.

U.S. President Richard Nixon feared an eventual Israeli confrontation with Moscow and sent Secretary of State William Rogers to intervene with a complex cease-fire proposal, which was accepted by Israel, Egypt, and Jordan in August 1970. This plan specified limits on the deployment of missiles and revived a year-old diplomatic initiative (the Rogers Plan) that insisted on an exchange of territory for peace on all fronts.

The Egyptians and Soviets soon violated the agreement by moving their missiles closer to the canal. In Jordan, Ḥussein's acceptance of the cease-fire ignited savage fighting between the Jordanian army and several PLO militia groups. As the battles intensified, Syria sent tanks to aid the Palestinians, but coordinated Israeli, American, and Jordanian military moves defeated the Syrians and expelled the PLO, whose forces sought refuge in Lebanon.

Meir's gamble had succeeded: Israel's willingness to risk confrontation had strengthened relations with the United States. Ḥussein's recovery of control in Jordan demoralized Palestinian resistance while securing Israel's eastern border. When Nasser died in September 1970, his successor, Anwar el-Sādāt, did not renew the fighting, seeking instead a partial Israeli pullback from the Suez Canal. Israel eventually rejected this idea, but the crisis had passed.

THE DECLINE OF LABOUR DOMINANCE

The Yom Kippur War. On October 6, 1973—the Jewish Day of Atonement, Yom Kippur—Egyptian and Syrian forces staged a surprise attack on Israeli forces situated on the Suez Canal and the Golan Heights. Israeli confidence in its early warning systems and air superiority was misplaced, and Egyptian missiles were soon taking a heavy toll of Israeli warplanes. The intensity of the Egyptian and Syrian assault, so unlike the situation in 1967, rapidly began to exhaust Israel's reserve stocks of munitions.

With Israel threatened by catastrophe, Prime Minister Meir turned to the United States for aid, while the Israeli general staff hastily improvised a battle strategy. Washington's reluctance to help Israel changed rapidly when the Soviet Union launched its own resupply effort to Egypt and Syria. President Nixon countered by establishing an emergency supply line to Israel. With reinforcements on the way, the IDF rapidly turned the tide. Israeli forces commanded by General Ariel Sharon crossed the Suez Canal and threatened to destroy the Egyptian Third Army. On the Golan, Israeli troops, at heavy cost, repulsed the Syrians and advanced to the edge of the Golan plateau on the road to Damascus. At this point, the United States, alarmed by Soviet threats of direct military intervention and on nuclear alert, secured a cease-fire in place.

Political and social repercussions of the war. Egypt's Sādāt persuaded U.S. Secretary of State Henry Kissinger that his country was ready to abandon both its Soviet and Syrian allies for a fresh start with the United States; only Washington, in Sādāt's view, could effectively influence Israel to return the Sinai without further bloodshed. Kissinger, supported by Nixon, successfully pressured Israel to end the war short of a complete Egyptian military defeat and then, through intensive travel between the various capitals—what soon was being called shuttle diplomacy—achieved disengagement agreements on both the Egyptian and Syrian fronts during 1974. This became known as the "step by step" process, which was intended to fulfill the intent of Security Council Resolution 242 that territory be exchanged for peace.

The "step by step" process

Golda Meir's government resigned in April 1974, exhausted and discredited by the war. Still, the Labour Party won a narrow election victory in June by selecting Yitzhak Rabin, hero of the 1967 war and former Israeli ambassador to Washington, to lead its list. Rabin predicted a period of "seven lean years" until the West, including the United States, would end its heavy dependence on Arab oil. He argued that Israel therefore needed to trade space for time, to coordinate closely with Washington, and to encourage Egypt's new pro-American policy.

Rabin reached a second disengagement agreement with Egypt in September 1975, but little progress was made with Syria. On what had been the "quiet" front—the West Bank and Gaza—the Labour government's preferred strategy of negotiations with the more amenable King Ḥussein (the so-called "Jordanian option") was threatened in October 1974, when an Arab summit conference in Rabat, Morocco, declared ʿArafāt's PLO to be the sole representative of the Palestinians. A year later, Rabin obtained secret assurances from Kissinger that the United States would not recognize the PLO as an entity representing the Palestinians unless that organization first ceased terrorism and recognized Israel's right to exist.

Meanwhile, the Gush Emunim movement on the West Bank gathered force after the Yom Kippur War and between 1974 and 1987 planted small communities near large Arab populations, greatly complicating Israeli policy and arousing international opposition. The secular Israeli government opposed such efforts but rarely used force to dislodge the settlers, who invoked Zionist rights to the homeland in their defense. Still, they numbered fewer than 4,000 when the opposition Likud government of Menachem Begin came to power in 1977.

Diplomatic impasse. The Yom Kippur War left the country in bad economic shape. An accelerating rate of inflation just before the conflict was suddenly combined with a stagnant economy; prices continued to rise even as demand and production fell. The international recession reduced demand for Israeli exports, and, for the first time in years, unemployment became a problem. On top of this, Israel became heavily indebted by arms purchases—partially offset by U.S. aid—and the country's international status suffered. One by one, most of Israel's carefully cultivated African friends broke relations under the threat of Arab oil sanctions, leaving the Jewish state alone with an equally isolated South Africa. Further complicating the situation for Israel, the UN General Assembly passed Resolution 3379 in 1975, which equated Zionism with racism, and the PLO gained increasing European and Asian support. Meanwhile, Rabin was losing political ground at home, harmed by infighting and corruption.

The "comprehensive approach"

A further blow to Rabin fell when he visited Washington in March 1977 to meet with the new American president, Jimmy Carter, who advocated a "comprehensive approach" to Middle East peace instead of the Kissinger step-by-step plan. Carter sought an international conference to resolve all of the major issues between Israel and the Arabs and advocated a "homeland" for the Palestinians. For Israelis, this notion (and its similarity to the wording of the Balfour Declaration) was a code word for a Palestinian state, and they hotly opposed it—not least because it also implied a leadership role for the PLO. Rabin, facing a major quarrel with the United States and beset by a personal scandal (his wife had maintained an illegal bank account in Washington from his days as ambassador), resigned in April.

ISRAEL UNDER LIKUD

To general surprise, the Likud Party, led by Menachem Begin, won the May 1977 election, inaugurating the first non-Labour-led government in Israel's history. Begin's campaign benefited immensely from Sephardic resentment over the patronizing attitude of the Labour establishment and its treatment of non-European Jewish immigrants as second-class citizens. The Sephardim supported Likud in large numbers. In addition, a new "clean government" party drew votes from Labour, and its leader, Yigael Yadin, an eminent archaeologist and a hero of the 1948 war, joined the cabinet. Also included in Begin's cabinet were Ezer Weizman, an air force commander in 1967 and architect of the Likud political strategy, and, to the shock of many Labourites, Moshe Dayan, who agreed to become foreign minister.

The beginning of the peace process. Begin strongly opposed any territorial compromise on the West Bank, which, like many Israelis, he felt to be an inalienable part of Israel—the historic Samaria and Judea. He also argued that Resolution 242 did not require withdrawal from this area. The new Israeli leader put off a crisis with Washington by discarding Rabin's notion of "coordination" and declared simply that Israel wanted to sit down with its neighbours to negotiate peace. Meanwhile, Begin inaugurated a policy to strengthen Israel's hold on the West Bank through an extensive settlement program overseen by Ariel Sharon, the minister of agriculture.

Extensive settlement program

Carter spent the summer in futile efforts to convene an international conference, finally approving a Soviet-American communiqué in October 1977 that was intended to stimulate diplomacy; instead, it outraged both Israel and the U.S. Congress, many of whose members condemned Carter for concessions to Moscow. These mishaps convinced Sādāt that American tactics were giving his erstwhile Soviet and Syrian allies a veto over any diplomacy, which could lead to a new war in which Egypt would likely pay the highest price. Secret negotiations were held between Israeli Foreign Minister Moshe Dayan and Sādāt's representatives, after which the Egyptian president surprised the world by flying to Israel in November 1977, where he and Begin addressed the Israeli Knesset.

The first invasion of Lebanon

The two leaders could not agree, however, on the details of a comprehensive peace, and the negotiations were complicated by events in Lebanon. Following its eviction from Jordan in 1971, the PLO had established itself there, exacerbating the volatile political situation in that country and contributing to its collapse into civil war in 1975. Both Israel and the United States had reluctantly consented to Syria's military intervention in Lebanon that same year, but the result was a partitioned state with the PLO dominating the south of the country, which was now a launching point for terror attacks against Israelis living in the Upper Galilee. In March 1978, Israel invaded Lebanon to drive the PLO away from the border but succeeded only partially in this goal before withdrawing from that country, under international pressure, in June.

Camp David. The faltering Egyptian-Israeli negotiations were finally rescued when Carter convened a summit at the presidential retreat of Camp David, Maryland, in September 1978. Out of these negotiations emerged the Camp David Accords. The accords were a framework for a comprehensive peace between Israel and the Arab states based on Resolution 242 and called for all of the parties to complete peace treaties under its principles. Rather than giving the Palestinians full independence, the accords offered them Begin's concept of autonomy, which provided for five years of limited Palestinian self-government to be followed by talks on final status between Israel and a Jordanian-Palestinian joint negotiating team.

The Camp David Accords earned Sādāt and Begin each a share of the 1978 Nobel Prize for Peace, but the subsequent peace process proved far more difficult than the parties expected. It took seven more months for Egypt and Israel to reach a final agreement, which was signed on March 26, 1979, and called for a three-year phased Israeli withdrawal from the Sinai, limited-force zones, a multinational observer force, full diplomatic relations between the two countries, and special provisions for Israeli access to the Sinai's oil fields. Israel's settlements in the Sinai were also evacuated, despite public Israeli protests.

Syria, Iraq, and the PLO were outraged by Egypt's actions and joined diplomatic forces to suspend Cairo from the Arab League and prevent any other Arab state from supporting the accords. Nearly all of the Arab states subsequently severed ties with Egypt. Jordan and the Palestinians refused to negotiate autonomy, and a three-year attempt by Israel, Egypt, and the United States to develop the plan on their own came to naught. Meanwhile, Begin refused to halt the building of new Jewish settlements in the West Bank and Gaza Strip.

A cold peace. Thus, "seven lean years" after the Yom Kippur War and three decades after independence, Israel had reached peace with Egypt, the Soviets were sidelined, and the Jewish state's alliance with the United States was consolidated. However, trouble loomed, as a civil war in Lebanon allowed an increasingly well-armed PLO to raid Israel's northern border. Israel had also begun to fear a military buildup in Iraq, especially its potential for producing nuclear weapons. Nor was the cabinet happy. Both Weizman and Dayan resigned from it, charging that the prime minister did not want to settle the Palestinian issue. The Begin government had also been much less successful in its domestic policies, and the economy, after a brief recovery in 1978–79, entered another inflationary spiral.

Begin's policies aroused strong international opposition but aided his victory over Shimon Peres in the June 1981 elections. His new government contained more Likud appointees, including Yitzhak Shamir as foreign minister and Ariel Sharon as defense minister. Then, on October 6, 1981, Sādāt was murdered by Muslim extremists. His successor, Hosnī Mubārak, reaffirmed the 1979 treaty but was prepared only for a "cold" peace with Israel, and few of the bright hopes for trade and tourism promised by the Camp David agreements materialized—even after Israel completed its withdrawal from the Sinai in April 1982.

War in Lebanon. Begin again turned to Lebanon, where he was determined to defeat the PLO. In July 1981, fearing an Israeli-Syrian clash in Lebanon, the United States had brokered an ambiguous cease-fire, during which the PLO continued to amass heavy arms. Cautioned by President Ronald Reagan's secretary of state, Alexander Haig, not to attack unless there was an "internationally recognized provocation," Begin ordered the bombing of PLO positions in June 1982 after members of a PLO splinter group attempted to assassinate Israel's ambassador to Britain. The PLO retaliated with a rocket barrage on Israel's northern border towns, whereupon Israel launched a new invasion of southern Lebanon. Israel's cabinet authorized a limited operation, and Begin made it clear that IDF troops were not to advance farther than 25 miles (40 km) beyond the Lebanese border. But Sharon had more ambitious plans. Even as Reagan's special envoy, Philip Habib, attempted to prevent an Israeli-Syrian clash, Israeli jets destroyed Syrian antiaircraft missiles in Lebanon. This strategic surprise attack was followed by a short but violent series of ground skirmishes and two days of aerial combat that cost Syria some 100 aircraft.

Sharon's gambit

Sharon sent the IDF toward Beirut and well beyond the mandated 25-mile limit. With the Syrians in retreat, Israeli troops besieged ʿArafāt and his remaining PLO units in the Lebanese capital. Israel's Maronite Christian allies, the Phalange Party, contrary to Sharon's expectations, did not act to secure the city as they had been expected to do, and a dangerous stalemate ensued. The pro-Israel Haig was forced from office, as a bewildered and angry Reagan, reinforced by U.S. Secretary of Defense Caspar Weinberger, sought an Israeli withdrawal. Habib, working under the direction of Haig's successor, George Shultz, managed to insert a multinational peacekeeping force in Lebanon that allowed ʿArafāt and a portion of his force to evacuate Beirut in August.

The Lebanese Christians, however, were not to benefit from the Israeli actions. Phalange leader Bashir Gemayel, the new president-elect, was assassinated by Syrian agents in September, and in the ensuing disorders, Israeli forces allowed the Phalangist militia into two Palestinian refugee camps, Sabra and Shatila, where they massacred hundreds of men, women, and children. The multinational force, withdrawn after ʿArafāt's departure, was reinserted.

Shortly before the massacres, President Reagan had announced a plan for Arab-Israeli peace that pointedly applied the Resolution 242 formula to the Palestinian issue. The plan was designed, in part, to appease Arab anger and to revive the Jordanian option, but it was rejected by an Arab summit and hotly opposed by an alarmed Begin. However, the embattled prime minister did not have much time left. An official Israeli inquiry condemned Sharon for negligence in the camp massacres, forcing him to resign. Grieving over Israeli losses and the operation's tragic outcome, Israelis mounted massive street demonstrations against the Begin government.

Under U.S. mediation, Israel and Lebanon reached a nonbelligerence agreement in May 1983, and Israeli troops withdrew from the Beirut area. An ailing Begin, devastated by his wife's death and the war's outcome, resigned in September and retreated into a reclusive retirement, dying in 1992. He was replaced by Yitzhak Shamir. The Syrians and their local allies later forced Lebanon to renounce the agreement with Israel.

THE NATIONAL UNITY GOVERNMENT

Labour outpolled the Likud in the 1984 election, but not by a margin sufficient to form a government. To rescue the economy and extricate Israel from its military entanglement in Lebanon, Labour and Likud formed a national unity government in September, giving the premiership to Peres for 25 months, at the end of which the premiership would go to Shamir, with the understanding that the other would take the position of deputy prime minister of foreign affairs. Notably, Rabin was to be defense minister for both men.

Under Peres, the Israelis began a phased withdrawal from Lebanon in June 1985, except for a security zone where an Israeli-sponsored Lebanese force waged intermittent warfare against a local Shīʿite militia, the Hezbollah, who enjoyed Iranian and Syrian patronage. For the first time, Israel began to reform its economic structures, which until then had been controlled by the state and the labour federation, Histadrut.

The Hezbollah

As stipulated by the rotation agreement, Shamir became prime minister in October 1986, with Peres as foreign minister. However, Peres was determined to regain the top spot through a diplomatic breakthrough and met secretly with Jordan's King Ḥussein. The two reached an understanding known as the London Agreement in April 1987, but the agreement's vague formulations did not command a majority of votes in the unity cabinet, and Shamir retained control.

Shamir continued the Begin policy of settling Jews throughout the West Bank, hoping to isolate the Arab towns and villages that might form the basis for a Palestinian state. Few Israelis responded to this initiative until Sharon, who returned to Shamir's cabinet as housing minister, began subsidizing residential communities that were within easy commuting distance of Jerusalem and Tel Aviv, where housing was scarce and expensive. By 1992 the Jewish population in the occupied territories was approaching 100,000.

The *intifāḍah*. The Begin and Shamir governments had gradually abandoned Moshe Dayan's old policy of leaving the Palestinians alone. By late 1987 the combined effects of settlement expansion, bureaucratic encroachment, land seizures, several years of economic stagnation, and the diplomatic stalemate had set the stage for an Arab rebellion in the West Bank and Gaza that quickly became known as the *intifāḍah* (Arabic: "shaking off"). This uprising was distinguished by widespread street violence in which children and teenagers battled Israeli troops with rocks and stones.

The Israeli military was caught by surprise and proved ill-equipped to deal with the revolt. A grinding contest of wills ensued that soon claimed many civilian casualties and altered the political landscape. In February 1988 Shamir invited Secretary of State Shultz to intervene, but he tried in vain to revive the diplomatic process. Meanwhile, King Ḥussein had finally abandoned his formal ambition to represent the Palestinians. Israel's international image was suffering as the media recorded scenes of Israeli soldiers beating young Palestinians in the street. Frequent closures of the areas also severely disrupted the Palestinian economy, and Israel began to replace Arab day labourers with immigrant workers from outside the region.

ʿArafāt recognizes Israel

The Israeli election in November 1988 gave Likud a slight majority. Shamir was still forming a government when in December ʿArafāt, speaking at a special UN meeting in Geneva, reiterated a declaration that he had made the previous month that he was ready to recognize Israel and suspend terrorism provided the Palestinians obtained a state. The United States promptly recognized the PLO and opened a dialogue with it.

The question of Palestinian autonomy. This stunning event led Shamir to form another national unity government, with Rabin again as defense minister and Peres as finance minister. Rabin was convinced that Israel needed a political initiative to end the *intifāḍah* and deflect the PLO. He persuaded Shamir to revive the Camp David-era autonomy plan, but this time it was stripped of its Jordanian component and aimed specifically at the Palestinians. Israel was also facing a new U.S. administration, led by President George Bush, that was determined to restrict Israeli settlement expansion. U.S. efforts to create an Israeli-Palestinian negotiation on autonomy, however, were rejected by Shamir, who insisted that the Palestinian negotiating team be drawn exclusively from residents of Gaza and the West Bank and not from Jerusalem or the

PLO. Peres thereupon resigned from the unity government, only to be outmaneuvered by Shamir, who formed a Likud-dominated coalition that excluded Labour. The prime minister decided to ride out the *intifāḍah* while concentrating on a sudden breakthrough with the Soviet Union: as part of Soviet leader Mikhail Gorbachev's reforms, a massive number of Soviet Jews were allowed to emigrate to Israel, the exodus continuing after the Russian Federation was created in the early 1990s.

The Gulf War and the Madrid Conference. The stalemated Arab-Israeli conflict was soon overshadowed by a crisis in the Persian Gulf, when the army of Iraqi leader Ṣaddām Ḥussein invaded Kuwait in August 1990. As the United States dispatched troops to Saudi Arabia and organized an international coalition against the Iraqi invasion, Ṣaddām attempted to stir up Arab antagonism against Israel. He found ready support among the Palestinians in Jordan and elsewhere, including an endorsement by PLO head ʿArafāt.

The United States greatly feared that its focus on Iraqi aggression would be diverted by Arab grievances against Israel, and, when the American-led coalition's attack was launched, Washington urged Israel not to respond to Iraqi provocations, even after Iraqi forces began missile attacks on Israeli cities. Accepting U.S. air-defense missiles, Israel held its fire while the coalition devastated the Jewish state's most dangerous Arab opponent. Meanwhile, the Persian Gulf states cut off their previously substantial financial support for the PLO.

Iraq's defeat and the decline of the Soviet Union in 1991 opened the way for fresh diplomatic initiatives. U.S. Secretary of State James Baker succeeded in convening an Arab-Israeli peace conference in Madrid in October 1991, the first direct official talks between Israel and its neighbours since the Camp David era. Three "tracks" were created under U.S. auspices that sought to achieve peace treaties between Israel and Jordan, Lebanon, and Syria; an interim Palestinian self-government for Gaza and the West Bank (the Palestinian team this time met the Israeli specifications); and European, Japanese, and Arab support for regional economic cooperation and arms control.

The talks stalled after a promising start. The Palestinians demanded statehood rather than autonomy, and Shamir was not interested in reaching quick agreements. The Israeli leader remained faithful to his strategy of outlasting the other side while continuing to construct Israeli settlements in the West Bank and Gaza. However, Shamir's policy was contested by the United States, and Bush refused Shamir's request for housing-loan guarantees to accommodate Russian immigrants unless Israel stopped expanding the settlements.

Continued settlement expansion

The Labour opposition, sensing an opportunity, put up Rabin as their candidate for prime minister in the elections of June 1992. He promised security but also flexibility, insisting that he would produce progress in the negotiations. He also proposed that less be spent on settlements and more on help for Russian immigrants. In a hard-fought election, the Labour Party won a narrow advantage.

THE RABIN GOVERNMENT

Rabin established greater control in this premiership than in his earlier one by keeping the defense portfolio to himself and appointing a negotiating team that reported to him rather than to Peres, his foreign minister. His coalition was delicately balanced between left and right and relied on a Sephardic religious party, Shas, to offset the strongly antireligious Meretz Party.

Rabin criticized the comprehensive approach implicit in the Madrid talks, concluding that the Palestinian-Israeli track held more promise for progress because both Israelis and Palestinians wanted to move beyond the status quo of the *intifāḍah.* To stimulate diplomacy and to patch up relations with the United States, he ordered a freeze on the construction of Israeli settlements in the occupied territories, which allowed the Bush administration to approve housing guarantees for Russian immigrants. (In fact, some previously planned construction continued in the territories, and the settler population grew from 100,000 to 135,000 during Rabin's term.)

Unexpectedly, the negotiations with Syria came to life first, but after an encouraging start they had deadlocked by the summer of 1993. Syria refused to specify what it meant by "full peace," a key Israeli requirement; and Israel refused to withdraw to the armistice lines as they were prior to the 1967 war, which would have effectively placed the border with Syria on Lake Tiberias (the Sea of Galilee), Israel's largest source of fresh water.

The Oslo Accords. Meanwhile, Peres had been nurturing a secret negotiating track with the Palestinians through Norwegian diplomacy. The PLO officials conducting the so-called unofficial discussions in Oslo, Norway, were far more flexible than the official non-PLO Palestinian delegation in Washington, and Rabin decided to gamble that ʿArafāt was the only Palestinian leader who could conceivably deliver peace. ʿArafāt also gambled. He was short of money after alienating his main financial backers during the Gulf War and faced challenges to his leadership from Islāmic groups, whose influence had grown significantly in the occupied territories during the *intifāḍah.* He accepted the idea of Palestinian autonomy in order at last to obtain a foothold in Palestine.

The Declaration of Principles and Cairo Agreement. In September 1993, Israel and the PLO signed the Declaration of Principles on Palestinian Self-Rule, the first agreement between the two sides and the initial document in what became generally known as the Oslo Accords. Both Rabin and ʿArafāt were happy to embrace U.S. President Bill Clinton on the White House lawn in September 1993, in support of their deal.

The Oslo Accords, in fact, comprised a series of agreements, the second of which, the Cairo Agreement on the Gaza Strip and Jericho, was signed in May 1994. This pact enacted the provisions set forth in the original declaration, which had endorsed a five-year interim self-rule for a Palestinian authority to be executed in two stages: first in Gaza and the city of Jericho and then, after an election, throughout the remaining areas under Israeli military rule. Talks on final status were to begin after three years. Issues such as borders, the return of refugees, the status of Jerusalem, and Jewish settlements in the occupied territories were reserved for final status talks. The PLO recognized Israel's right to exist, renounced terrorism, and agreed to change the portions of its charter that called for Israel's destruction. Israel recognized the PLO as the sole representative of the Palestinian people.

Israel recognizes the PLO

The accords embodied two basic sets of exchanges. First, Israel would shed responsibility for the Palestinian population while retaining strategic control of the territory. The Palestinians would be rid of Israeli military rule and gain self-government, potentially leading to statehood. Second, ʿArafāt's disavowal of violence and his pledge to fight terrorism would improve Israel's security.

Challenges to peace. Rabin's decision aroused enormous opposition from the Likud and most settlers, although the majority of Israelis at first strongly supported him. In October 1994, Jordan also signed a comprehensive peace treaty with Israel, and many other Arab states, including the smaller Persian Gulf emirates, began to discard the old taboos about contact with the Jewish state.

Not all Palestinians, however, favoured ʿArafāt's course. The Islāmic group Ḥamās, which was especially strong in Gaza, violently opposed the Oslo Accords and launched a series of terror attacks on Israeli civilians, killing scores between 1993 and 1997. Rabin retaliated with border closures that prevented tens of thousands of Palestinian workers living in the occupied territories from commuting to jobs in Israel. Some Israelis sought revenge, such as the murder in February 1994 of some three dozen Arabs at prayer in Hebron's Tomb of the Patriarchs by an Israeli settler. Despite Israeli protests, ʿArafāt sought to co-opt rather than repress Ḥamās.

Economic boom. Peace diplomacy bolstered what had already been a period of strong economic expansion in Israel. Austerity during the 1980s had wrung out bad debt and inefficiency at considerable cost. Many kibbutzim, deprived of cheap credit and a subsidized water supply, had either failed or shifted from agriculture to light industry. Koor Industries Ltd., Histadrut's industrial holding company, had itself fallen on hard times and defaulted on a number of loans before it was restructured. The Israeli government still controlled half of the economy, but the earlier socialist ideology, once the mainstay of Israeli politics, was clearly on the wane.

Socialism on the wane

In the late 1980s the Israeli economy was buoyed by the influx of highly skilled Russian immigrants, a competitive high-technology sector, and the country's proximity to the European market. In the period 1990–95, Israel's rate of economic growth exceeded 5 percent annually, unemployment was cut nearly in half, and the annual inflation rate dropped from double to single digits. Foreign investment turned from a trickle into a flood, as Israeli exports to Asia also registered large increases and the Arab boycott eased.

Oslo II and Rabin's assassination. In September 1995, Rabin, ʿArafāt, and Peres, all newly named winners of the Nobel Prize for Peace, assembled again on the White House lawn to sign the Interim Agreement on the West Bank and Gaza Strip (often called Oslo II). This detailed and long-delayed agreement established a schedule for Israeli withdrawals from the Palestinian population centres and created a complex system of zones that were divided between areas fully controlled by the Palestinians, those under Palestinian civil authority but Israeli military control, and those exclusively under Israeli control.

Although Oslo I had received strong parliamentary support, Oslo II was ratified by only one vote in the Knesset, signaling a significant loss of support for Rabin. Many Israelis were angry over ʿArafāt's erratic cooperation on security, and others, especially the Likud—now led by Israel's former ambassador to the UN, Benjamin Netanyahu—opposed withdrawals or further dealings with ʿArafāt. Meanwhile, the Sephardic Shas Party had left the coalition in protest over the indictment of its parliamentary leader for fraud. Bereft of his coalition's balance, Rabin had to depend on the vote of the Israeli Arab members of the Knesset. He was also battered by demands from the Meretz Party and from Conservative and Reform Jews in the United States to loosen the Orthodox religious monopoly established in the early years of the state.

Shortly after Oslo II was passed in the Knesset, Rabin decided on a public campaign to rally his supporters, and it was following the first such rally in Tel Aviv in November 1995 that he was assassinated by a Jewish religious fanatic. Israelis were horrified, and, after a funeral attended by many international leaders, including Arabs, a round of soul-searching and recriminations began. Popular Israeli support for the peace process surged, and, with the Likud on the defensive, Shimon Peres, Rabin's successor as prime minister, proceeded with Oslo II. By early 1996 nearly all of the Palestinians were under self-rule; Israeli forces, though withdrawn from the major towns except Hebron, still controlled most of the occupied territories. In January, ʿArafāt was elected president of the Palestinian Authority.

A NEW POLITICAL LANDSCAPE

Peres had hoped to capitalize on sympathy for Rabin and chose to hold early elections in 1996. His campaign was quickly upset by a series of Ḥamās suicide attacks against civilians that shocked and angered Israelis. Further, the prime minister lacked Rabin's security credentials and had been an outspoken advocate of partnership with ʿArafāt. Peres reacted to Hezbollah attacks along the Lebanese border by ordering a massive artillery bombardment of southern Lebanon that mistakenly hit a United Nations outpost sheltering hundreds of civilians, which damaged his standing among Israeli Arabs. In addition, Labour had lost popularity among religious voters because of its alliance with Meretz.

The Netanyahu premiership. The May elections were the first held under a new law that allowed separate ballots for prime minister and the Knesset, which was designed to reduce the ability of smaller parties to exact concessions when governments were formed. However, this law had the opposite result: it created a quasi-presidential regime that still depended on an increasingly fractured parliament. Peres narrowly lost to Netanyahu, who promised that he would be tougher on ʿArafāt than Peres had been. In the Knesset, however, both Labour and Likud unexpectedly lost ground, while the smaller parties, especially the religious bloc, gained large numbers of seats. An ethnic Russian party, Yisrael B'Aliyah, also won seats. The growth of the Shas Party and the emergence of a Russian ethnic bloc offered compelling evidence not only of their grievances against previous governments but of the failure of the major political parties to integrate these constituencies.

Growth of ethnic parties

Netanyahu promised to accelerate economic reforms, especially the sale of state-owned businesses, but he was quickly confronted by labour union opposition, a slowing economy, and a large budget deficit. He had been a severe critic of the Oslo Accords but, after Rabin's murder, had promised to fix the agreements by insisting on Palestinian "reciprocity" (*i.e.,* strict adherence to the terms). Nonetheless, Netanyahu could not bring himself to meet ʿArafāt until September 1996 and raised doubts over his willingness to proceed with the promised Israeli withdrawal from Hebron and other unfinished aspects of Oslo II.

Crisis in the peace process. ʿArafāt had finally acted against Ḥamās in the spring and summer of 1996 but soon concluded that Netanyahu was not engaging in serious negotiating. In late September, Israel's surprise opening of an archaeological tunnel exit along the Western Wall in Jerusalem allowed ʿArafāt to play on Palestinian fears that the tunnel was a threat to the al-Aqṣā Mosque. Another violent round of protests ensued, in which Palestinian police fired on Israeli soldiers, killing more than a dozen, while some 60 Arabs died. Clinton, fearing the end of the peace process, called an emergency summit in the White House. Netanyahu refused to close the tunnel, but he pledged an accelerated effort to negotiate an agreement on the West Bank city of Hebron, which was reached in January 1997.

Netanyahu's cabinet narrowly approved the Hebron Agreement. Part of the price for this action became clear when Israel began constructing a long-planned but often delayed Jewish neighbourhood on the outskirts of Jerusalem, which would effectively cut off the Arab villages on the eastern side of the city from the rest of the West Bank. ʿArafāt held his protest of this project until the cabinet's decision on the first of three projected Israeli withdrawals in March. When these withdrawals turned out to be far less significant than ʿArafāt had anticipated, the stage was set for another round of protests that quickly escalated into violence. Meanwhile, ʿArafāt released Ḥamās activists from Palestinian jails and suspended security cooperation with Israel. Netanyahu, fearing that his cabinet would not support any more interim steps, argued that Israel and the Palestinians should begin intensive negotiations to determine the final status of the territories. This proposal was quickly rejected by both ʿArafāt and the United States.

Release of Ḥamās activists

The Wye River Memorandum. The breakdown of direct Israeli-Palestinian negotiation at high levels led the United States to intervene again in early 1998 to end the stalemate. Both sides met in rural Maryland in October and produced the Wye River Memorandum. The new agreement restored old Israeli promises (such as the opening of a Palestinian airport and a safe passage route between Gaza and the West Bank) for old Palestinian promises (such as publicly renouncing the PLO Charter's anti-Israel provisions, collecting unauthorized arms, and implementing antiterrorist actions), but its novelty consisted of linking phased Israeli withdrawals to Palestinian actions and greatly enlarging the role of the United States as an active participant in both monitoring and judging the performance of the parties. Wye promised to put two-fifths of the West Bank under partial or total Palestinian control.

Netanyahu returned from Wye to face growing political trouble. The Bank of Israel had been using high interest rates in a dramatic effort to reduce Israeli inflation. While the policy succeeded overall, it also precipitated a recession and rising unemployment, which hit hardest in the poorer sectors of society—notably the largely Sephardic residents of the development towns in the south. Concurrently, the government's budget had been reduced, which hampered the prime minister's ability to satisfy the demands of the various coalition members. In early 1999, after a legislative defeat on the budget, Netanyahu called for early elections and soon suspended the Wye agreement.

The Barak gamble. The May 1999 Israeli election produced an even more fractured Knesset than the one three years earlier. Whereas in 1992, under the old, purely parliamentary system, the two largest parties had between them won 76 of the 120 seats, by 1999 they could command only 45. Labour, renamed One Israel in coalition with two small parties, had the most seats in the Knesset, while Likud, beset by infighting and a stalled economy, was second. The real surprise of the election was the sudden growth of Shas, which now commanded the third-largest number of seats.

Negotiations intensify. Labour's candidate for prime minister was a retired general, Ehud Barak, who triumphed over Netanyahu by a decisive margin. Barak had promised a renewed drive for peace, economic growth, and resistance to religious demands. He assembled a broad coalition in the Knesset and set about reviving the peace process with both the Palestinians and Syrians.

Like his predecessor, Barak wanted definitive talks with ʿArafāt about the final status of the territories before vacating much more land, but he encountered heavy resistance to his plans from both foreign and domestic sources. The Palestinians would not agree to abandon the third and final troop withdrawal promised under Oslo II; ʿArafāt put off his planned declaration of a Palestinian state but insisted on maximum American intervention and sought the most territory he could recover before the final negotiations. However, he did agree to Barak's deadline of February 15, 2000, to reach a framework agreement, which was to be preceded by another withdrawal. These new arrangements were incorporated in the Wye II agreement, reached in September 1999. None of the deadlines was met.

Talks with Syria

As the prime minister expected and ʿArafāt feared, the Syrians suddenly signaled their desire to negotiate in early December. Barak himself traveled to the United States to negotiate with his counterpart, Syrian Foreign Minister Fārūq al-Sharʿ, under Clinton's patronage. A second session in early January 2000, however, ended when Syrian President Ḥafiz Assad broke off the talks, raising the old demand that Israel agree to a return to the borders that existed between the two countries before the Six-Day War as a precondition to further negotiations.

By early March, however, progress again seemed possible on the Syrian front. Assad agreed to a summit with President Clinton at Geneva, but to U.S. and Israeli surprise he yet again insisted on Syria's right to its pre-1967 positions on Lake Tiberias. Neither Barak nor a majority of Israelis would agree to this demand. Barak then carried out his campaign promise to withdraw Israeli troops from Lebanon, even without an agreement with Syria, to a border demarcated by the UN.

Barak's willingness to concede to Arab demands and his mishandling of his coalition destroyed his Knesset majority in June. Nonetheless, he decided to attend Clinton's summit at Camp David in July. This last-ditch effort to reach an agreement between ʿArafāt and Barak had been resisted by the Palestinian leader, who stated ahead of time that he could not concede Palestinian rights. This proved to be the case. Barak's unexpected willingness to share Jerusalem with the Palestinians was not reciprocated by ʿArafāt, who on this—as on the issue of the return of refugees—refused to compromise, demanding nothing less than full Palestinian sovereignty in East Jerusalem.

The second intifāḍah. ʿArafāt returned home from the summit to Palestinian and popular Arab acclaim. He had said "no" to both Israel and the United States. In contrast, Barak's political support evaporated, and, as he struggled to survive, a new blow fell when the Palestinians erupted in violence following Likud leader Sharon's visit to the Temple Mount in September. Rioting by Israeli Arabs further disturbed the situation. As international efforts to restore peace failed, cameras recorded the death of a 12-year-old Arab boy by gunfire in Gaza, and not long thereafter two Israeli soldiers were lynched in the West Bank. By spring 2001, hundreds had been killed, most of them Palestinians. President Clinton made one last attempt to bridge the gap in December 2000, but neither side accepted his "parameters."

The return of Sharon

The failure of the Camp David summit and the outbreak of what came to be known as the al-Aqṣā *intifāḍah* convinced a majority of Israelis that they lacked a partner in ʿArafāt to end the conflict. Barak paid the political price, losing the premiership to Sharon in elections held in February 2001 by nearly 25 percent of the vote. (Ha.Si.)

BIBLIOGRAPHY

Physical and human geography. General information is available in *Country Review: Israel* (semiannual), published by CountryWatch.com. Studies in historical geography include RUTH KARK (ed.), *The Land That Became Israel* (1989); and ELISHA EFRAT, *Geography and Politics in Israel Since 1967* (1988). LEO PICARD, *Structure and Evolution of Palestine: With Comparative Notes on Neighbouring Countries* (1943, reprinted 1959), is a comprehensive geologic study. MICHAEL ZOHARY, *Plant Life of Palestine: Israel and Jordan* (1962), is a fundamental work; and F.S. BODENHEIMER, *Animal Life in Palestine: An Introduction to the Problems of Animal Ecology and Zoogeography* (1938), is a classic work on Israel's fauna. Settlements are discussed in AHARON KELLERMAN, *Society and Settlement: Jewish Land of Israel in the Twentieth Century* (1993).

Economic studies include YAIR AHARONI, *The Israeli Economy: Dreams and Realities* (1991); and YAKIR PLESSNER, *The Political Economy of Israel: From Ideology to Stagnation* (1994). Administrative and political issues are explored by DON PERETZ and GIDEON DORON, *The Government and Politics of Israel*, 3rd ed. (1997); ALAN DOWTY, *The Jewish State: A Century Later* (1998); YAEL YISHAI, *Land of Paradoxes: Interest Politics in Israel* (1991); and GREGORY S. MAHLER (ed.), *Israel* (2000). ADAM GARFINKLE, *Politics and Society in Modern Israel*, 2nd ed. (2000), discusses Israel's domestic and foreign policy. ASHER ARIAN (ALAN ARIAN), *The Second Republic: Politics in Israel* (1998; originally published in Hebrew, 1997), is based on unique electoral demographic data.

CALVIN GOLDSCHEIDER, *Israel's Changing Society: Population, Ethnicity, and Development* (1996), is a useful work. Sociological studies include: ELIEZER BEN-RAFAEL and STEPHEN SHAROT, *Ethnicity, Religion, and Class in Israeli Society* (1991); and GABRIEL BEN-DOR (ed.), *Israel in Transition* (1998). Israeli culture is analyzed by AVISHAI MARGALIT, *Views in Review: Politics and Culture in the State of the Jews* (1998). (Ru.A.S.)

History. General surveys of the region's prehistory are KATHLEEN M. KENYON, *Archaeology in the Holy Land*, 4th ed. (1979, reprinted 1985); and ARCHAEOLOGICAL INSTITUTE OF AMERICA, *Archaeological Discoveries in the Holy Land* (1967). Works describing the Zionist movement and the establishment of Israel include NAHUM SOKOLOW, *History of Zionism, 1600–1918*, 2 vol. (1919, reprinted 2 vol. in 1, 1969); ALBERT M. HYAMSON, *Palestine Under the Mandate, 1920–1948* (1950, reprinted 1976); and BARNET LITVINOFF, *To the House of Their Fathers: A History of Zionism* (1965). Valuable interpretations are contained in PETER Y. MEDDING, *The Founding of Israeli Democracy, 1948–1967* (1990); and LAURENCE J. SILBERSTEIN, *New Perspectives on Israeli History: The Early Years of the State* (1991).

The literature on Israel's history since 1948 includes T.G. FRASER, *The Arab-Israeli Conflict* (1995); BERNARD REICH *et al.* (eds.), *An Historical Encyclopedia of the Arab-Israeli Conflict* (1996); and HOWARD M. SACHAR, *A History of Israel: From the Rise of Zionism to Our Times*, 2nd ed., rev. and updated (1996). BENNY MORRIS, *Righteous Victory: A History of the Zionist-Arab Conflict, 1881–1999* (1999), is a good example of the controversial "new historians" school; and MARTIN GILBERT, *Atlas of the Arab-Israeli Conflict*, 6th ed. (also published as *The Dent Atlas of the Arab-Israeli Conflict*, 1993), provides geographic context.

Palestinian autonomy is discussed in HARVEY SICHERMAN, *Palestinian Self-Government (Autonomy): Its Past and Its Future* (1991; also published as *Palestinian Autonomy, Self-Government & Peace*, 1993); and an analysis of events within the Israel-Lebanon-Syria triangle is MOSHE MAʿOZ, *Syria and Israel: From War to Peacemaking* (1995). Literature concerning Israel's relationship with the United States includes WILLIAM B. QUANDT, *Peace Process: American Diplomacy and the Arab-Israeli Conflict since 1967* (1993); and BERNARD REICH, *Securing the Covenant: United States-Israel Relations After the Cold War* (1995). DOV S. ZAKHEIM, *Flight of the Lavi* (1996), opens the window on the vital U.S.-Israeli military relationship. Finally, good accounts of Israel's peace diplomacy include ITAMAR RABINOVITCH, *The Brink of Peace: The Israeli-Syrian Negotiations* (1998); and URI SAVIR, *The Process: 1,100 Days That Changed the Middle East* (1998). (Ha.Si.)

Istanbul

Istanbul (formerly Constantinople, ancient Byzantium, Turkish İstanbul) is the largest city and seaport of Turkey. It was formerly the capital of the Byzantine Empire, of the Ottoman Empire, and—until 1923—of the Turkish Republic.

The old, walled city of Istanbul stands on a triangular peninsula between Europe and Asia. Sometimes as a bridge, sometimes as a barrier, Istanbul for more than 2,500 years has stood between conflicting surges of religion, culture, and imperial power. For most of those years it was one of the most coveted cities in the world.

This article is divided into the following sections:

Physical and human geography

THE LANDSCAPE

The city site. The old city contains about nine square miles (23 square kilometres), but the present municipal boundaries stretch for more than 98 square miles, including areas on both sides of the Bosporus and the Sea of Marmara. The original peninsular city has seven hills requisite for Constantine's "new Rome." Six are crests of a long ridge above the Golden Horn; the other is a solitary eminence in the southwest corner.

The so-called three seas

By long tradition, the waters washing the peninsula are called "the three seas": they are the Golden Horn, the Bosporus, and the Sea of Marmara. The Golden Horn is a deep, drowned valley about 4½ miles (seven kilometres) long. Early inhabitants saw it as being shaped like a deer horn, but modern Turks call it the Haliç (Canal). The Bosporus (Boğaziçi) is the channel connecting the Black Sea (Kara Deniz) to the Mediterranean (Ak Deniz) by way of the Sea of Marmara (Marmara Deniz) and the straits of the Dardanelles. The narrow Golden Horn separates the old city of Stamboul to the south from the "new" city of Beyoğlu to the north; the broader Bosporus divides European Istanbul from the city's districts on the Asian shore—Üsküdar (ancient Chrysopolis) and Kadiköy (ancient Chalcedon).

Climate. The prevailing northeast wind, or *poyraz,* comes from the Black Sea, giving way at times during the winter to an icy blast from the Balkans known as the *karayel,* or "black veil," capable of freezing the Golden Horn and even the Bosporus. The *lodos,* or southwest wind, can raise storms on the Sea of Marmara.

The city plan. Stamboul is still a walled city. The land walls, which isolate the peninsula from the mainland, were breached only once, by the cannon of Ottoman sultan Mehmed II (the Conqueror) in 1453, at the spot since called Cannon Gate (Top Kapısı). The walls are four and a half miles long and consist of a double line of ramparts—the inner built in 413, the outer in 447—protected by a moat. The higher inner wall is about 30 feet (9 metres) high and 16 feet thick and is studded with 60-foot towers about 180 feet apart. Of 92 turrets originally raised on the outer wall, 56 are still standing.

The sea walls were built in 439. Only short sections of their 30-foot-high masonry still remain along the Golden Horn. Intact, these walls had 110 towers and 14 gates. The walls along the Sea of Marmara, which stretch about five miles from Seraglio Point, curving around the bottom of the peninsula to join the land walls, had 188 towers; they were, however, only about 20 feet high, because the Marmara currents provided good protection against enemy landings. Most of these walls still stand.

Within the city walls are the seven hills, their summits flattened through the ages, their slopes still steep and toilsome. Geographers number them from the seaward tip of the peninsula, proceeding inland along the Golden Horn, the last hill standing alone where the land walls reach the Sea of Marmara.

The Golden Horn

The Galata and Atatürk bridges cross the Golden Horn to Beyoğlu. Each day before dawn their centre spans are swung open to allow passage to seagoing ships. The shores of the Horn, served by water buses, are a jumble of docks, warehouses, factories, and occasional historical ruins. Ferries to the Asian side of Istanbul leave from under the Galata Bridge. Istanbul's Bosporus Bridge (completed in 1973), with a main span of 3,524 feet, is one of the world's longest suspension bridges. A second span crossing the Bosporus, the Fatih Sultan Mehmed Bridge, was completed in 1988.

Beyoğlu, considered to be "modern Istanbul," remains, as it has been since the 10th century, the foreign quarter. Warfare and fires have left standing only a few structures that were built earlier than the 19th century.

The approach from the Golden Horn is steep, and a funicular railway runs between the Galata waterfront and the Pera Plateau. On the heights are the big hotels and restaurants, the travel bureaus, theatres, the opera house, the consulates, and many Turkish government offices.

From the 10th century, Galata was an enclave for foreign traders—principally the Genoese—who enjoyed extraterritorial privileges behind their walls. After the Ottomans took the city in 1453, all foreigners who were not citizens of the empire were restricted to this quarter. Eventually Galata became too crowded, so that the tide of building moved higher up the slope to the open country of Pera.

Remains of Constantine's city

Architecture. *Byzantine monuments.* Nothing remains of the Byzantium that Constantine chose as the site of New Rome, and little is left of the city he built there. Constantine's column, the Burnt Column (Çemberlitaş), a shaft of porphyry drums bound by metal laurel leaves, still stands near the Mosque of Nûruosman, but there is no proof that any building in the city dates from his period. He completed the Hippodrome that Septimus Severus had restored, but it was enlarged and rebuilt by his successors until the 5th century. Only its curved end remains, with three columns along the central spina—an obelisk removed from Egypt by the Roman emperor Theodosius I, a masonry obelisk of Constantine VII Porphyrogenitus (AD 905–959), and a Delphic column formed by three entwined serpents (now headless) cast after the Battle of Plataea, when the Greeks defeated the Persians in 479 BC.

Of the myriad columns that decorated Constantinople, there remain standing only the base of the column of the emperor Arcadius (reigned 383–408) in the Cerrahpaşa quarter; a column of the emperor Marcian (reigned 450–457) that the Turks call Kıztaşı; the Column of the Virgin, in the Fatih quarter; and, in the grounds of the Topkapı Palace, a perfectly preserved Corinthian column thought to be from the reign of another emperor, Claudius II Gothicus (268–270).

Spanning the valley between the third and fourth hills is the two-story limestone aqueduct built in 366 by the emperor Valens. Some of the enormous open-water cisterns of the Byzantine epoch now serve as market gardens. The closed cisterns, of which there are more than 80 remaining, include one of the most beautiful and mysterious struc-

Istanbul and (inset) its metropolitan area.

tures of Istanbul, the Basilican Cistern, or Yerebatan Sarayı (underground palace), near the Hagia Sophia; its 336 columns rise from the still waters to a vaulted roof.

The Golden Gate is a triumphal arch from about 390 and was built into the defenses of Theodosius II, near the junction of the land and sea walls. The marble-clad bases of its two large towers still stand, and three arches decorated with columns stretch between them.

The only well-preserved example of Byzantine palace architecture is the shell of a three-story rectangular building of limestone and brick, laid in patterns and stripes. Dating from about 1300, it is called the Palace of Constantine (Tekfur Sarayı) and is attached to the land walls not far from the Golden Horn.

The largest legacy from the capital of the vanished empire is constituted by 25 Byzantine churches. Many of these are

The Golden Horn, Istanbul. The two mosques in the background are (centre left) Hagia Sophia and (centre right) the Mosque of Süleyman.
Ara Gule

still in use—as mosques. The largest of the churches is considered one of the great buildings of the world. This is the Hagia Sophia (*q.v.*), whose name means "Divine Wisdom." Its contemporary and neighbour, St. Irene, was dedicated to "Divine Peace." Many art historians deem the dome (105 feet in diameter) of Hagia Sophia to be the most beautiful in the world. The Church of SS. Sergius and Bacchus was erected by Justinian between 527 and 536 as a thank offering. It is also called the Mosque of Küçük Ayasofya (Little Sophia) and can be considered an architectural parent of Justinian's reconstruction of the Hagia Sophia. The Church of St. Saviour in Chora (now called the Mosque of Kari) is near the Adrianople Gate. It was restored in the 11th century and remodelled in the 14th; the building is now a museum renowned for its 14th-century mosaics, marbles, and frescoes.

The Hagia Sophia

Turkish monuments. When the Ottomans took possession of Constantinople, they covered the spines of the seven hills with domes and minarets, changing the character of the city. The most imposing of their mosques were constructed from the mid-15th to the mid-16th century, and the greatest of the architects all bore the name of Sinan. They were Atik Sinan (the Elder), Sinan of Balıkesir, and Mimar Koca Sinan (Great Architect Sinan). Although the building was deeply influenced by the Persian-born traditions of the Seljuq Turks, the style was blended with prevailing Hellenic and Byzantine traditions of the city. Mimar Koca Sinan's masterpiece—and his burial place—is the Mosque of Süleyman (1550–57), inspired by, but not copied from, the Hagia Sophia. It ranks as another of the world's great buildings. Probably the most popularly known of all the mosques in Istanbul is the Blue Mosque, that of the Mosque of Ahmed I (Ottoman sultan from 1603 to 1617), which has six minarets instead of the customary four.

The Sinan architectural tradition

The mosques of the 18th century and later show the deleterious effects of importing European architects and craftsmen, who produced Baroque Islāmic architecture (such as the Mosque of the Fatih, rebuilt between 1767 and 1771) and even Neoclassical styles, as in the Dolmabahçe Mosque of 1853, now the Naval Museum.

There are more than 400 fountains in Istanbul. Some simply flow from wall niches, but others, erected as public philanthropies, are pavilions. The most magnificent of these was built by the sultan Ahmed III in 1728, behind the apse of Hagia Sophia. It is square, with marble walls and bronze gratings, a mixture of the Turkish with the Western Rococo style.

To the north of it, toward the Golden Horn and occupying the whole tip of the promontory, is the sultan's Seraglio (Topkapı Palace), enclosed in a fortified wall. It was begun in 1462 by Mehmed II and served as the residence of the sultans until the beginning of the 19th century. The Seraglio consists mostly of small buildings grouped around three courts. The most significant buildings are the Çinili Kiosk (Pavilion of Tiles) built in 1472, the Audience Chamber (Arzodası), the Hirkaiserif, a sanctuary containing relics of the Prophet Muḥammad, and the elegant Baghdad Kiosk commemorating the capture of Baghdad in 1638. The Seraglio houses the sultan's treasure and has important collections of manuscripts, china, armour, and textiles. After the abandonment of the Old Seraglio, the sultans built for themselves palaces along the Bosporus, such as the Beylerbey Palace (1865), the lavish Dolmabahçe Palace (1853), the Çeragan Palace built in 1874 and burned in 1910, and the Yıldız Palace, which was the residence of Abdülhamid II, Ottoman sultan from 1876 to 1909.

Topkapı Palace

THE PEOPLE

Istanbul, like other major cities in the region, attracts an increasing number of migrants from the countryside. These migrants have contributed to the growth of shantytowns called *gecekondu* (literally "set down by night") that have no sanitation facilities and limited access to power and water. The Christian and Jewish minorities continue to shrink both in percentage of the whole and in overall numbers. Kurds now constitute the largest ethnic minority in the city.

THE ECONOMY

Industry. Istanbul is Turkey's largest port and is the hub of its industry. Textiles, flour milling, tobacco processing, cement, and glass are the city's principal manufactures. Tourism is a growing source of income for Istanbul.

Transportation. There is rail service along the walls of the old city, and Haydarpaşa station, on the Asian side of the city, is the starting point of the Baghdad Railway. Maritime services include many forms of transport, from harbour dinghies and small ferries to international liners. Yeşilköy Airport is about 17 miles to the west of the city. Buses provide internal urban transportation, and the ferries range as far as the Kızıl Adalar (Princes Islands), several hours sailing to the south. Atatürk International Airport is close to the city's centre and provides both overseas and domestic service.

ADMINISTRATION AND SOCIAL CONDITIONS

Government. The mayor, appointed by the president of the republic, serves as prefect of Istanbul city and governor of Istanbul *il* (province). The municipality, which was organized by Constantine as 14 districts in imitation of Rome, is now divided into 12 circumscriptions (*kazas*), each governed by a *kaymakam*.

Public utilities. While Istanbul has a chlorinated and filtered water supply and sewage disposal system, these facilities are not sufficient to meet the increased need created by the influx of rural migrants to the city. Water supply is a problem particularly in the summer when rivers run dry; at this season tap water is likely to flow only sporadically, except in luxury hotels. Electric power supplies have been increased to help promote industrial expansion.

Health. Most of the health services of Istanbul *il* (province) are concentrated in the municipality. There are more than 70 hospitals, about half of which are public.

Istanbul's universities

Education. The first University of Istanbul was founded in 425 by Theodosius II and was succeeded by Istanbul University (İstanbul Üniversitesi), founded in 1453. The university now includes faculties of letters, science, law, medicine, and forestry located in the former Seraskerat (war ministry) between the Great Bazaar and the Mosque of Süleyman. There is also a technical university on the Galata side of the Horn as well as an Academy of Fine Arts and schools of technology, commerce, and economics. Foreign educational institutions include the American Robert College for boys (founded in 1863) and the American College for girls (founded in 1871), both on the Bosporus.

CULTURAL LIFE

The Palais de la Culture d'Istanbul is an important centre for the arts. Facilities include a concert hall, art gallery, and two theatres. It is the home of the Istanbul Municipal Symphony Orchestra and the Istanbul City Opera. The municipal theatre operates several playhouses, and there are many theatre companies.

A large number of learned societies and research institutes are headquartered in the city, including the Geographical Institute (Coğrafya Enstitüsü), German and French archaeological institutes, and the Turkish Folklore Society (Türk Halk Bilgisi Derneği). There is a nuclear research centre at Küçük Çekmece.

There are many public and private libraries. The small, specialized Köprülü Library (1677) has books from early Ottoman presses and handwritten works more than 1,000 years old. Many of the city's mosques, palaces, and monuments, as mentioned earlier, contain museums; other museums include the Archaeological Museums of Istanbul (Istanbul Arkeoloji Müzeleri), the Museum of Turkish and Islāmic Art (Türk ve Islam Eserleri Müzesi), and the Museum of the Janissaries (Türkiye Askeri Müzesi).

The Hippodrome is now a public garden; there are also numerous other public parks. A unique feature of the city is its market gardens; these kitchen gardens are associated with the open cisterns that formed early Constantinople's water-supply system. The cisterns have been partially built over and are called Çukur Bostan (Hollow Gardens).

Football (soccer) is a popular sport, and Istanbul has three stadiums—Mithatpaşa, Fenerbahçe, and the indoor Spor ve Sergi Sarayı. There are facilities for tennis, fencing, mountain climbing, riding, golf, and water sports. Florya and Ataköy are popular beaches on the Sea of Marmara.

History

THE EARLY PERIOD

The founding of Byzantium

Byzantium. Byzantium was one of the many colonies founded from the end of the 8th century onward along the coasts of the Bosporus and the Black Sea by Greek settlers from the cities of Miletus and Megara.

The Persian king Darius I took the settlement in 512 BC; it slipped from Persian grasp during the Ionian revolt of 496, only to be retaken by the Persians. In 478 an Athenian fleet captured the city, which then became a rich and important member of the Delian League. As Athenian power waned during the Peloponnesian War, Byzantines acknowledged Spartan overlordship. Although Alcibiades besieged and retook the city, Sparta reasserted its domination after defeating Athens in 405 BC.

In 343 BC Byzantium joined the Second Athenian League, throwing off the siege of Philip II of Macedon three years later. A free city under Rome, it gradually fell under imperial control and briefly lost its freedom under the emperor Vespasian. When, in AD 196, it sided with the usurper Pescennius Niger, the Roman emperor Septimus Severus massacred the populace, razed the walls, and annexed the remains to the city of Perinthus (or Heraclea, modern Marmaraereğlisi), in Turkey.

Subsequently, Septimus Severus rebuilt the city on the same spot but on a grander scale. In the subsequent civil wars and rebellions that broke out sporadically in the Roman Empire, Byzantium remained untouched until the arrival of the emperor Constantine I—the first Roman ruler to adopt Christianity. Overcoming the army of the rival emperor, Licinius, at nearby Chrysopolis, on September 18, 324, Constantine became head of the whole Roman Empire, east and west. He decided to make Byzantium his capital.

The New Rome

Constantinople. Constantine's new city walls tripled the size of Byzantium, which now contained imperial buildings, such as the completed Hippodrome begun by Severus, a huge palace, legislative halls, several imposing churches, and streets decorated with multitudes of statues taken from rival cities. In addition to other attractions of the capital, free bread and citizenship were bestowed on those settlers who would fill the empty reaches beyond the old walls. There was a welcome for Christians, a tolerance of pagan beliefs, and benevolence toward Jews.

Constantinople was also an ecclesiastical centre. In 381 it became the seat of a patriarch who was second only to the bishop of Rome; the patriarch of Constantinople is still the nominal head of the Orthodox church. Constantine inaugurated the first ecumenical councils; the first six were held in or near Constantinople.

The pressure of population pressing from within, and the barbarian threat from without, prompted the building of walls further inland at the hilt of the peninsula. These new walls of the early 5th century, built in the reign of Theodosius II, are those that stand today.

In the reign of Justinian I (527–565) medieval Constantinople attained its zenith. At the beginning of this reign the population is estimated to have been about 500,000. In 532 a large part of the city was burned and many of the population killed in the course of the repression of the Nika Insurrection, an uprising of the Hippodrome factions. The rebuilding of the ravaged city gave Justinian the opportunity to engage in a program of magnificent construction, of which many buildings still remain.

In 542 the city was struck by a plague that is said to have killed three out of every five inhabitants; the decline of Constantinople dates from this catastrophe. Not only the capital but the whole empire languished, and slow recovery was not visible until the 9th century. During this period the city was frequently besieged—by the Persians and Avars (626), the Arabs (674 to 678 and again from 717 to 718), the Bulgars (813 and 913), the Russians (860, 941, and 1043), and by a wandering Turkic people, the Pechenegs (1090–91). All were unsuccessful.

In 1082 the Venetians were allotted quarters in the city itself (there was an earlier cantonment for foreign traders at Galata) with special trading privileges. They were later joined by Pisans, Amalfitans, Genoese, and others. These Italian groups soon obtained a stranglehold over the city's foreign trade—a monopoly that was finally broken by a massacre of Italians. Not for some time were Italian traders permitted once more to settle in Galata.

Crusader rule

In 1203 the armies of the Fourth Crusade, deflected from their objective in the Holy Land, appeared before Constantinople—ostensibly to restore the legitimate Byzantine emperor, Isaac II. Although the city fell, it remained under its own government for a year. On April 13, 1204, however, the crusaders burst into the city to sack it. After a general massacre, the pillage went on for years. The crusading knights installed one of themselves, Baldwin of Flanders, as emperor, and the Venetians—prime instigators of the crusade—took control of the church. While the Latins divided the rest of the realm among themselves, the Byzantines entrenched themselves across the Bosporus at Nicaea (now İznik) and at Epirus (now northwestern Greece). The period of Latin rule (1204 to 1261) was the most disastrous in the history of Constantinople. Even the bronze statues were melted down for coin; everything of value was taken. Sacred relics were torn from the sanctuaries and dispatched to religious establishments in western Europe.

In 1261 Constantinople was retaken by Michael VIII Palaeologus, Greek emperor of Nicaea. For the next two centuries the shrunken Byzantine Empire, threatened both from the West and by the rising power of the Ottoman Turks in Asia Minor, led a precarious existence. Some construction was carried out in the late 13th and early 14th centuries, but thereafter the city was in decay, full of ruins and tracts of deserted ground, contrasting with the prosperous condition of Galata across the Golden Horn, which had been granted to the Genoese by the Byzantine ruler Michael VIII. When the Turks crossed into Europe in the mid-14th century, the fate of Constantinople was sealed. The inevitable end was retarded by the defeat of the Turks at the hands of Timur (Tamerlane) in 1402; but in 1422 the Ottoman sultan of Turkey, Murad II, laid siege to Constantinople. This attempt failed, only to be repeated 30 years later. In 1452 another Ottoman sultan, Mehmed II, proceeded to blockade the Bosporus by the erection of a strong fortress at its narrowest point; this fortress, called Rumeli Hisarı, still forms one of the principal landmarks of the straits. The siege of the city began in April 1453. The Turks had not only overwhelming numerical superiority but also cannon that breached the ancient walls. The Golden Horn was protected by a chain, but the sultan succeeded in hauling his fleet by land from the Bosporus into the Golden Horn. The final assault was made on May 29, and, in spite of the desperate resistance of the inhabitants aided by the Genoese, the city fell. The last Byzantine emperor, Constantine XI Palaeologus, was killed in battle. For three days the city was abandoned to pillage and massacre, after which order was restored by the sultan.

Capture by the Turks

CENTURIES OF GROWTH

When Constantinople was captured, it was almost deserted. Mehmed II began to repeople it by transferring to it populations from other conquered areas such as the Peloponnese, Salonika (modern Thessaloníki), and the Greek islands. By about 1480 the population rose to between 60,000 and 70,000. Hagia Sophia and other Byzantine churches were transformed into mosques. The Greek patriarchate was retained, but moved to the Church of the Pammakaristos Virgin (Mosque of Fethiye), later to find a permanent home in the Fener (Phanar) quarter. The sultan built the Old Seraglio (Eski Saray), now destroyed, on the site occupied at present by the university, and a little later the Topkapı Palace (Seraglio), which is still in existence; he also built the Eyüp Mosque at the head of the Golden Horn and the Mosque of the Fatih on the site of the Basilica of the Holy Apostles. The capital of the Ottoman Empire was transferred to Constantinople from Adrianople (Edirne) in 1457.

After Mehmed II, Istanbul underwent a long period of peaceful growth, interrupted only by natural disasters—earthquakes, fires, and pestilences. The sultans and their ministers devoted themselves to the building of fountains, mosques, palaces, and charitable foundations so that the aspect of the city was soon completely transformed. The most brilliant period of Turkish construction coincides with the reign of the Ottoman ruler Süleyman the Magnificent (1520–66).

The next major change in the history of Istanbul occurred at the beginning of the 19th century, when dismemberment of the Ottoman Empire was approaching. This period was known as the era of internal reforms (Tanzimat). The reforms were accompanied by serious disturbances, such as the massacre of the Janissaries in the Hippodrome (1826). With the triumph of the progressive Ottoman sultan Mahmud II over the conservative opposition, the westernization of Istanbul started apace. There was an ever-growing influx of European visitors who, since the 1830s, could reach Istanbul by steamship. The first bridge across the Golden Horn was built in 1838. In 1839 the Ottoman sultan Abdülmecid I issued a charter guaranteeing to all his subjects, whatever their religion, the security of their lives and fortunes. The process of westernization was further accelerated by the Crimean War (1853–56) and the quartering of British and French troops in Istanbul. The latter part of the 19th and the beginning of the 20th centuries were marked by the introduction of various public services: the European railroad extending to Istanbul was begun in the early 1870s. The underground tunnel joining Galata to Pera was completed in 1873; a regular water supply for Istanbul and the settlements on the European side of the Bosporus was brought from Lake Terkos on the Black Sea coast (29 miles from the city) by the French company, La Compagnie des Eaux, after 1885; electric lighting was introduced in 1912 and electric street cars and telephones in 1913 and 1914. An adequate sewage system had to wait until 1925 and later. (B.E.)

Westernization of the city

MODERN ISTANBUL

In the first quarter of the 20th century, there were various disruptions marking the death of the Ottoman Empire and the birth of modern Turkey. In 1908 the city was occupied by the army of the Young Turks who deposed the hated sultan Abdülhamid II. During the Balkan Wars (1912–13) Istanbul was nearly captured by the Bulgarians. Throughout World War I the city was under blockade. After the conclusion of the Armistice (1918) it was placed under British, French, and Italian occupation that lasted until 1923. The Greco-Turkish War in Asia Minor, as well as the Russian Revolution, brought thousands of refugees to Istanbul. With the victory of the Nationalists under Mustafa Kemal Atatürk, the sultanate was abolished, and the last Ottoman sultan, Mehmed VI, fled from Istanbul (1922). After the signing of the Treaty of Lausanne, Istanbul was evacuated by the Allies (October 2, 1923), and Ankara was chosen as the capital of Turkey (October 13, 1923). On October 29, the Turkish Republic was proclaimed. Because of Turkey's neutrality during most of World War II, Istanbul suffered no damage, although a German invasion was feared after the Balkans had been conquered by the Axis.

In the period following World War II, the size and population of Istanbul increased dramatically, as vast numbers of rural residents moved to the city in search of employment. This nearly 10-fold increase in the city's population during the second half of the 20th century placed enormous strains on Istanbul's infrastructure, and, in a pattern typical of large Middle Eastern cities, overcrowding, pollution, and insufficient city services became major social problems. Likewise, in a region prone to violent seismic activity, the proliferation of substandard and unregistered construction contributed greatly to high death tolls during earthquakes, as in an August 1999 tremor centred near Istanbul in which more than 15,000 persons died.

These developments took place against the backdrop of a city whose profile was being rapidly altered by an explosion in the use of automobiles. Large tracts of the city were demolished or cleared to make way for modern highways, which further contributed to urban sprawl, and major projects have been undertaken to connect the Asian and European sides of the city by road and rail.

BIBLIOGRAPHY

Antiquities: A. VAN MILLINGEN, *Byzantine Constantinople* (1899); PHILIP SHERRARD, *Constantinople* (1965); PHILIP GRIERSON, *The Tombs and Obits of the Byzantine Emperors, 337–1042* (1962); BERNARD LEWIS, *Istanbul and the Civilization of the Ottoman Empire* (1963); DEAN A. MILLER, *Imperial Constantinople* (1969); MICHAEL MACLAGAN, *The City of Constantinople* (1968).

Churches: T. WHITTEMORE, *The Mosaics of St. Sophia at Istanbul,* 4 vol. (1933–52); ROBERT MARK and AHMET Ş. ÇAKMAK (ed.), *Hagia Sophia from the Age of Justinian to the Present* (1992); PAUL ATKINS UNDERWOOD, *The Kariye Djami* (1966).

Contemporary descriptions: ROBERT BOULANGER, *Istanbul et ses environs* (1957; Eng. trans. 1960); PETER MAYNE, *Istanbul* (1967); CHRIS HELLIER, *Splendors of Istanbul: Houses and Palaces along the Bosporus* (1993). (B.E./Ed.)

Italian Literature

Vernacular Italian literature had its beginnings in the 13th century. Until that time nearly all literary work was written in Latin, was predominantly practical in nature, and was produced by writers trained in ecclesiastical schools. In literary quality and variety it fell short of the standard set by France. Only small fragments of Italian vernacular verse before the end of the 12th century have been found (although a number of legal documents were written in the vernacular), and surviving 12th- and 13th-century verse reflects French influence.

This article traces the development of Italian literature from early French-inspired works to the diverse writings of present-day authors.

The article is divided into the following sections:

EARLY VERNACULAR LITERATURE

Franco-Italian literature. French prose and verse romances were popular in Italy from the 12th to the 14th century. Stories from the Carolingian and Arthurian cycles, together with free adaptations from the classics, were read by the literate, while French minstrels recited verse in public places throughout northern Italy. By the 13th century a "Franco-Italian" literature had developed; Italians copied French stories, often adapting and extending various episodes and sometimes creating new romances in French about characters from French works. In this literature, though the language used was French, the writers often unconsciously introduced elements from their own dialects, according to their varying knowledge of French. Writers of important prose works such as Martino da Canale and Brunetto Latini—who wrote *La Cronique des Veniciens* (1275; "Chronicle of the Venetians") and *Li Livres dou trésor* (*c.* 1260; "The Books of the Treasure"), respectively—were much better acquainted with French, while poets such as Sordello of Mantua wrote lyrics in Provençal revealing an exact knowledge of the language and of Provençal versification. Provençal love lyrics were, in fact, as popular as the French romances, and Italian writers carefully studied anthologies of the verse.

The origins of the vernacular lyric

The Sicilian school. In the cultured environment of the Sicilian court of the Holy Roman emperor Frederick II, who ruled the Sicilian kingdom from 1208 to 1250, lyrics modeled on Provençal forms and themes were written in the vernacular. Poets were careful to eliminate narrowly local elements from their language and used words characteristic of the troubadour tradition. Poetry was considered an escape from serious matters of life, and it is significant that it was the love poetry of Provence—and not the political poetry—that was imitated by the Sicilian school.

The Tuscan poets. Sicilian poetry continued to be written after the death of Frederick II, but the centre of literary activity moved to Tuscany, where interest in the Sicilian lyric had led to several imitations by Guittone d'Arezzo and his followers. Although Guittone experimented with elaborate verse forms, his language mingled dialect elements with Latinisms and Provençalisms and had none of the beauty of the southern school.

The new style. While Guittone and his followers were still writing, a new development appeared in love poetry, marked by a concern for precise and sincere expression and a new, serious treatment of love. It became customary to speak of a new school of poets of the *dolce stil novo,* or *nuovo* ("sweet new style"), an expression used by Dante Alighieri in *La divina commedia* (*Purgatorio,* Canto XXIV, line 27) in a passage where he emphasized delicacy of expression suited to the subject of love. The major *stil novo* poets were Guido Guinizelli of Bologna, Guido Cavalcanti, Dante (in his poems in *La vita nuova*), and Cino da Pistoia, together with the lesser poets Lapo Gianni, Gianni Alfani, and Dino Frescobaldi.

These poets seem to have been influenced by each other's work. Guido Guinizelli was best known for his canzone, or poem, beginning "Al cor gentil ripara sempre amore" ("Love always finds shelter in the gentle heart"), which posed the question of the relationship between love of woman and love of God. His poetry was immediately appreciated by Cavalcanti, a serious and extremely talented lyric poet. Most of his poems were tragic and denied the ennobling effect of love suggested by Guinizelli. Dante greatly admired Cavalcanti, but his concept of love, inspired by his love for Beatrice, who died young (in 1290), had much more in common with Guinizelli's. Dante's *Vita nuova* (*c.* 1293; *The New Life*) is the story of his love in poems linked by a framework of eloquent prose: God is the "root" of Beatrice, and she is able to mediate God's truth and love and inspire love of God—but her death is necessary for her lover to reach a state of purification. Cino da Pistoia used the vocabulary of the *stilnovisti,* as these poets were called, in an original and lighthearted way. A comparison of their language with the earlier Tuscan poets reveals extensive refinement of Tuscan dialects. Purely local characteristics were removed, and the standard literary language of Italy had been created.

Dante's *Vita nuova*

Comic verse. Giocoso, or comic, verse was a complete contrast to serious love poetry. The language was often deliberately unrefined, colloquial, and sometimes obscene, in keeping with the themes dealt with in the poetry. This kind of verse belongs to a European tradition, owing something to the satirical Goliard poets of the 12th and 13th centuries, who wrote Latin verses in praise of pleasure or in vituperation of women or personal enemies or the church. The comic poets—whose usual verse form was the sonnet—were all cultured men. The earliest of them was Rustico di Filippo, who produced both courtly love poetry and coarse, sometimes obscene verse of the "realistic" kind. The best known and most versatile was Cecco Angiolieri, whose love poetry often skillfully parodied the *stil novo* writers and whose favourite subject was his father's meanness. Folgore di San Gimignano was known for his sonnets (following Latin models) on the worldly pleasures he considered as suitable to different months and different days of the week.

Religious poetry. The beautiful and famous *Cantico di frate sole* (*c.* 1225; "Canticle of the Sun") of St. Francis of Assisi was one of the earliest Italian poems. It was written in rhythmical prose that used assonance in place of rhyme and was in the Umbrian dialect; in it God is praised through all the things of his creation. It is probable that St. Francis also composed a musical accompaniment, and after his death the *lauda* became a common form of religious song used by the confraternities of lay people who gathered to sing the praises of God and the saints and to recall the life and Passion of Christ. The one real poet of the *laude* was Jacopone da Todi, a Franciscan and a mystic. His *laudi* were mostly concerned with the theme of spiritual poverty. Though Jacopone did not write with conscious art, some of the *laudi* were very fine.

The pessimism of northern Italian poetry

In northern Italy religious poetry was mainly moralistic and pervaded by a pessimism rooted in heretical ideas derived from Manichaeism, which saw the world and the body as being evil and under Satan's control. Giacomino da Verona, a Franciscan, author of *De Jerusalem celesti* (*c.* 1250; "On Heavenly Jerusalem") and *De Babilonia civitate infernali* (*c.* 1250; "On the Infernal Babylonian State"), was the liveliest and most imaginative of this group.

Prose. Literary vernacular prose began in the 13th century, though Latin continued to be used for writings on theology, philosophy, law, politics, and science.

The founder of Italian rhetorical prose style, Guido Faba, a rhetorician, illustrated his teaching by examples of prose styles in Bolognese. Guittone d'Arezzo, his most notable follower in epistolography, tended toward an extravagant style. In contrast with Guittone's style is the clear scientific prose of Ristoro d'Arezzo's *Della composizione del mondo* (1282; "On the Composition of the World") and the simple narrative of the Florentine collection of tales *Il novellino* (written in the late 13th century, published in 1525 as *Le ciento novelle antike; Il Novellino, the Hundred Old Tales*). The masterpiece of 13th-century prose is Dante's *Vita nuova.* Though not yet completely at ease in vernacular prose, Dante combined simplicity with great delicacy and a poetic power that derived from the mysterious depth beneath certain key words.

THE 14TH CENTURY

The literature of 14th-century Italy dominated Europe for centuries to follow and may be regarded as the starting point of the Renaissance. Three names stand out: Dante, Petrarch, and Boccaccio.

Dante. Dante Alighieri is one of the most important and influential names in all European literature, but it was only after his exile (1302) that he set out to write more ambitious works. *Il convivio* (*c.* 1304–07; "The Banquet"), revealing his detailed knowledge of scholastic philosophy, was the first great example of a treatise in vernacular prose: its language avoided the ingenuousness of popular writers and the artificiality of the translators from Latin. *De vulgari eloquentia* ("On Vernacular Eloquence"), written about the same time, contained the first theoretical discussion and definition of the Italian literary language. Both these works remained unfinished. In a later doctrinal work, *De monarchia* (written *c.* 1313; Eng. trans., *De Monarchia*), Dante expounded his political theories, which demanded the coordination of the two medieval powers, pope and emperor.

La divina commedia

Dante's genius found its fullest development in *La divina commedia* (written *c.* 1310–21; *The Divine Comedy*), an allegorical poem in terza rima (stanzas of three lines of 11 syllables each, rhyming *aba, bcb, cdc,* etc.), the literary masterpiece of the Middle Ages and one of the greatest products of any human mind. The central allegory of the poem was essentially medieval, taking the form of a journey through the world beyond the grave, with, as guides, the Roman poet Virgil and Beatrice, who symbolize reason and faith, respectively. The poem is divided into three *cantiche,* or narrative poems: *Inferno, Purgatorio,* and *Paradiso.* They are subdivided into various schematically symbolic areas, and each contains 33 cantos, with one canto as an overall prologue. Dante, through his experiences and encounters on the journey, gains understanding of the gradations of damnation, expiation, and beatitude, and the climax of the poem is his momentary vision of God. The greatness of the poem lies in complex imaginative power of construction, inexhaustible wealth of poetry, and continuing significance of spiritual meanings.

Petrarch. The intellectual interests of Petrarch (Francesco Petrarca, died 1374) were literary rather than philosophical; his political views were more realistic than Dante's and his poetic technique more elaborate though less powerful. Petrarch's influence on literature was enormous and lasting—stretching through the Italian humanists of the following century to poets and scholars throughout western Europe. He rejected medieval Scholasticism and took as his models the classical Latin authors and the Church Fathers. This convergence of interests is apparent in his philosophical and religious works. Humanist ideals inspired his Latin poem *Africa* (begun *c.* 1338) and his historical works, but the autobiographical *Secretum meum* (written 1342–58; *Petrarch's Secret*) is most important for a full understanding of his conflicting ideals. The *Rime,* or *Canzoniere*—a collection of sonnets, songs, six-line verses, ballads, and madrigals, dating from 1330 until his death—gave these ideals poetic expression. Although this collection of vernacular poems intended to tell the story of his love for Laura, it was in fact an analysis and evocation not of present love but of the passion that he had overcome. The main element of this poetry was therefore in the elaboration of its art, even if it always reflected the genuine spiritual conflicts exposed in the *Secretum.* In addition to the *Canzoniere* Petrarch wrote a vernacular allegorical poem, *Trionfi* (1351–74; *Triumphs*), in the medieval tradition, but it lacked the moral and poetical inspiration of Dante's great poem.

The literary phenomenon known as Petrarchism developed rapidly within the poet's lifetime and continued to grow during the following three centuries, deeply influencing the literatures of Italy, Spain, France, and England. His followers did not merely imitate but accepted his practice of strict literary discipline and his forms, including that for the sonnet—without which the European literary Renaissance would be unthinkable.

Boccaccio. Boccaccio's writings were purely literary, without any ideological implications. His first romance, *Il filocolo* (*c.* 1336; "Love's Labour"), derived from the French romance *Floire et Blancheflor,* was little more than literary experiment. Inability to write on an epic scale was evident in his two narrative poems *Il filostrato* (*c.* 1338; "Frustrated by Love") and *Teseida* (*c.* 1340; *The Book of Theseus*), while his *Ameto* (1341–42), a novel written in prose and verse, and his *Fiammetta* (*c.* 1343; *Amorous Fiammetta*), a prose novel, showed the influence of classical literature on the formation of his style. The *Decameron* (1348–53), a prose collection of 100 stories divided into 10 "days," was Boccaccio's most mature and important work. Its treatment of contemporary urban society ranged from the humorous to the tragic. Stylistically the most perfect example of Italian classical prose, it had enormous influence on Renaissance literature.

The *Decameron*

Boccaccio shared the humanist interests of his age, as shown in his Latin epistles and treatises. An admirer of Dante, he also wrote a *Trattatello in laude di Dante* (written *c.* 1360; "Treatise in Praise of Dante"; Eng. trans., *The Life of Dante*) and a commentary on the first 17 cantos of the *Inferno*. He contributed to allegorical poetry with *L'amorosa visione* (written 1342–43).

Popular literature and romances. During the second half of the 14th century, Florence remained a centre of culture, but its literature developed a more popular character. The best known representative of this development was Antonio Pucci (died 1388), whose vast production included the *Centiloquio* (written after 1348; "One Hundred Tales"), a versification of Giovanni Villani's *Cronica*. Florentine narrative literature was represented by the *Pecorone* (written *c.* 1378; "Dullard"), stories by Ser Giovanni Fiorentino after a pattern established by Boccaccio, and Franco Sacchetti's *Trecentonovelle* (written *c.* 1390; "300 Short Stories"), which provide colourful and lively descriptions of people and places.

The recasting of the Carolingian and Arthurian cycles continued along lines established during the 13th century. Compilations in prose and verse became commoner, and Franco-Venetian literature gained in literary value. Epic legends were turned into romantic stories, which appealed more to their audiences in town squares and other public places. Novels by Andrea da Barberino, *cantari* with legendary subjects by Antonio Pucci, and the anonymous *Pulzella gaia, Bel Gherardino, Donna del Vergiù,* and *Liombruno* were written in the popular style consistent with their practical aim.

Religious and historical literature. The most important author of religious literature was Jacopo Passavanti, whose *Specchio di vera penitenza* ("The Mirror of True Penitence") was a collection of sermons preached in 1354. Less polished, but of greater literary value, were translations of legends collected in the anonymous *Fioretti di San Francesco* (*The Little Flowers of St. Francis of Assisi*).

The Florentine chroniclers

Vernacular historiography of this period could be described as popular literature, with Florence as its main centre, whose two principal chroniclers were Dino Compagni and Giovanni Villani. Compagni wrote his chronicle between 1310 and 1312, after having taken part in the political struggles of his town; his dramatic account of the episodes and the liveliness of his prose made it the most original work of medieval historiography. Villani's *Cronica* in 12 books, written from 1308 to 1348, was less personal; it followed the medieval tradition by beginning with the building of the Tower of Babel and included many legends. The last six books, which cover the period from Charles of Anjou's Italian expedition (1265) to the author's own time, are of importance to historians. His prose lacked the dramatic power of Compagni's, but his work may be described as the greatest achievement of Italian vernacular historiography during the Middle Ages.

From Boccaccio's death to about the middle of the 15th century, Italian poetry suffered a decline. The following period was to be characterized by critical and philological activity rather than by original creative work.

THE RENAISSANCE

The age of humanism. The European Renaissance had really begun in 14th-century Italy with Petrarch and Boccaccio. The 15th century, devoid as it was of major poetic works, was nevertheless of very great importance because it was the century in which a new vision of human life, embracing a different conception of man and life as well as more modern principles of ethics and politics, gradually found its expression. This was the result, on the one hand, of the rediscovery of classical antiquity, and, on the other, of political conditions quite different from those of previous centuries. With regard to the second point, nearly all Italian princes competed with each other in the 15th century to promote culture by patronizing research, offering hospitality and financial support to literary men of the time, and founding libraries. As a consequence, their courts became centres of research and discussion, thus making possible the great cultural revival of the period. The most notable courts were that of Florence, under Lorenzo de' Medici, "the Magnificent"; that of Naples, under the Aragonese kings; that of Milan, first under the Visconti and later the Sforza family; and finally the papal court at Rome, which gave protection and support to a large number of Italian and Byzantine scholars. As for the first point, the search for lost manuscripts of ancient authors, begun in the second half of the previous century, led to an extraordinary revival of interest in classical antiquity: in particular, much research was devoted to Plato and Greek philosophy in general, a fact that was to have profound influence on the thinking of the Renaissance as a whole.

Italian humanist view of man

By and large, the new culture of the 15th century was a revaluation of man. Humanism opposed the medieval view of man as a being with relatively little value and extolled him as the centre of the universe, the power of his soul as linking the temporal and the spiritual, and earthly life as a realm in which the soul applies its powers. These concepts, which mainly resulted from the new interest in Plato, were the subject of many treatises, the most important of which were Giannozzo Manetti's *De dignitate et excellentia hominis* (completed in 1452; *On the Dignity of Man*) and Pico della Mirandola's *Oratio de hominis dignitate* (written 1486; *Oration on the Dignity of Man*). The humanist vision evolved during this period condemned many religious tenets of the Middle Ages still widely prevalent: monastic ideals, for example, were attacked by Leonardo Bruni, Lorenzo Valla, and Poggio Bracciolini. Forthright though these attacks were, humanism was not essentially anti-Christian, for it generally remained faithful to Christian beliefs, and the papal court itself regarded humanism as a force to be assimilated rather than defeated.

In the first half of the century humanists, with their enthusiasm for Latin and Greek literature, had a disdain for the Italian vernacular. Their poetic production, inspired by classical models and written mostly in Latin or occasionally Greek, was abundant but of little value. Writing in a dead language and closely following a culture to which they had enslaved themselves, they rarely showed originality as poets. Among the few notable exceptions are Giovanni Pontano, Michele Marullo Tarcaniota, Angelo Ambrogini Poliziano (Politian), and Jacopo Sannazzaro. These poets sometimes succeeded in creating sincere poetry in which the conventional themes of 15th-century lyrics were expressed with new, original intimacy and fervour.

The rise of vernacular literature. Toward the middle of the 15th century Italian began to oust Latin as the literary language. The Certame Coronario, a public poetry competition held in Florence in 1441 with the intention of proving that the spoken language was in no way inferior to Latin, marked a definite change. In the second half of the century there were a number of works of merit inspired either by Carolingian legends or by the new humanist culture.

The chivalrous epic of Carolingian legends, which had degenerated into clichés, was given a new lease on life by two poets of very different temperament and education: Matteo Maria Boiardo, whose *Orlando innamorato* (1483; "Orlando in Love") reflected past chivalrous ideals as well as contemporary standards of conduct and popular passions; and Luigi Pulci, whose *Morgante,* published before 1480, was pervaded by a new bourgeois and popular morality.

The new ideals of the humanists were most complete in Angelo Ambrogini Poliziano (Politian), Jacopo Sannazzaro, and Leon Battista Alberti, three outstanding figures who combined a wide knowledge of classical antiquity with a personal and often profound inspiration. Politian's most important Italian work was *Stanze per la giostra* (1475–78; "Stanzas for the Joust"), which created a mythical world in which concepts of classical origin were relived in a new way. The same could be said of Sannazzaro's *Arcadia* (1504), a largely autobiographical work in verse and prose that remained widely influential up to the 18th century. A more balanced view of contemporary reality was given in Alberti's literary works, which presented a gloomy picture of human life, dominated by man's wickedness and the whims of fortune. As for Lorenzo de' Medici, patron of many men of letters, he himself had a vast poetic output, though this is more notable for documentary than for literary value.

Pietro Bembo's view of language

Pietro Bembo of Venice published his *Prose della volgar lingua* ("Writings in the Vulgar Tongue") in 1525. In this work, which was one of the first Italian grammars, Bembo demanded an Italian literary language based on 14th-century Tuscan models, particularly Petrarch and Boccaccio. He was opposed by those who thought that a literary language should be based on existing linguistic developments, particularly by Gian Giorgio Trissino, who developed Dante's theories on Italian as a literary language. In practice the problem was both linguistic and stylistic, and there were in the first half of the 16th century a great number of other contributors to the question, but Bembo's theories did finally triumph in the second part of the century. This was due to a large extent to the activities of the Florentine Accademia della Crusca, and this more scientific approach to the language question resulted in the academy's first edition of an Italian dictionary in 1612.

During the first decades of the 16th century, treatises on poetry were still composed according to humanist ideas and the teachings of the Roman Augustan poet Horace. It was only after 1536, when the original classical Greek text of Aristotle's *Poetics* was first published, that a gradual development became apparent in aesthetic theory. The traditional principle of imitation was now better analyzed, emphasis being given to the imitation of classical authors rather than to that of nature. The three unities of tragedy (time, space, action) were among the rhetorical rules then reestablished. The classical conception of poetry as a product of imagination supported by reason was at the basis of 16th-century rhetoric, and it was this conception of poetry, revived by Italian literature, that triumphed in France, Spain, and England during the following century.

Political, historical, biographical, and moral literature. Niccolò Machiavelli's works reflected Renaissance thought in its most original aspects, particularly in the objective analysis of human nature. Machiavelli has been described as the founder of a new political science: politics divorced from ethics. His own political experience was at the basis of his ideas, which he developed according to such general principles as the concepts of *virtù* ("power") and *fortuna* ("chance"). He considered *virtù* to be power with a practical aim that should struggle against *fortuna*, which represented the forces of violence and irresponsibility. His famous treatise *Il principe* (*The Prince*), composed in 1513, revealed the author's prophetic attitude, based on observation of contemporary political affairs. Its description of a model ruler became a code for the wielding of absolute power throughout Europe for two centuries. Machiavelli's *Discorsi sopra la prima deca di Tito Livio* (*c.* 1513–21; *Discourse on the First Ten Books of Titus Livius*), showed the same realistic attitude: public utility was placed above all other considerations, and political virtue was distinguished from moral virtue. His seven books on *Dell'arte della guerra* (1521; *The Art of War*), concerning the creation of a modern army, were more technical, while his historical works, including the *Istorie fiorentine* (1520–25; *Florentine History*), exemplified theories expounded in his treatises. Machiavelli also holds a place in the history of imaginative literature, above all for *La Mandragola* (1518), one of the outstanding comedies of the century.

The Prince

Although more of an individualist and pragmatist than Machiavelli, Francesco Guicciardini was the only 16th-century historian who could be placed within the framework of the political theories he constructed. He drew attention to the self-interest of those involved in political actions and made Machiavelli's theories appear idealistic by contrast. One of Guicciardini's main works, his *Ricordi* (1512–30; "Remarks"), has a place among the most original political writings of the century. Guicciardini was also the first, in his *Storia d'Italia* (1537–40), to compose a truly national history of Italy, setting it in a European context and attempting an impartial analysis of cause and effect.

Giorgio Vasari's *Vite de' più eccellenti architetti, pittori et scultori italiani da Cimabue insino a' tempi nostri* (1568; *Lives of the Painters, Sculptors, and Architects*) not only contained more than 200 biographies but also was the first proper critical and historical appraisal of Italian art. The autobiography of the sculptor and goldsmith Benvenuto Cellini (written 1558–66, published 1728) was remarkable for its spontaneity and its use of popular Florentine language.

The highest moral aspirations of the Renaissance are expressed in Baldassare Castiglione's *Cortegiano* (published 1528; *The Courtier*), which deals with the perfect courtier, the noble lady, and the relationship between courtier and prince. It became one of the most influential books of the century. Giovanni della Casa was the author of another famous treatise, the *Galateo* (*c.* 1551–54; "Manners"; Eng. trans., *Galateo*), a book on courtesy in which the author's witty mind and the refinement of contemporary Italian society found full expression. The life of the period was also vividly reflected in the work of Pietro Aretino, who was called "the scourge of princes" by Ariosto. His *Ragionamenti* (1534–36; "Discussions") were written in a spontaneous style and showed a sensuous and unscrupulous nature.

Castiglione's *Cortegiano*

Poetry. Lyric poetry in the 16th century was dominated by the model of Petrarch mainly because of the acceptance of the Renaissance theory of imitation and the teaching of Bembo. Almost all the principal writers of the century wrote lyric poems in the manner of Petrarch. Some originality was to be found in Della Casa's poems, and Galeazzo di Tarsia stood out from contemporary poets by virtue of a vigorous style. Also worthy of note are the passionate sonnets of the Paduan poet Gaspara Stampa and those of Michelangelo.

The tradition of autobiographical, humorous, and satirical verse was kept alive during the 16th century, when it reached some real stature with Francesco Berni, whose burlesque poems, mostly dealing with indecent or trivial subjects, showed his stylistic skill. Didactic poetry, already cultivated by humanist writers, was also continued during this period, chiefly by Giovanni Rucellai, who recast in *Le api* (1539; "The Bees") the fourth book of the Roman poet Virgil's *Georgics*, and Luigi Alamanni, in six books about rustic life called *La coltivazione* (1546).

The most refined expression of the classical taste of the Renaissance was to be found in Ludovico Ariosto's *Orlando furioso* (1516; Eng. trans., *Orlando Furioso*), which embodied many episodes derived from popular medieval and early Renaissance epics; but the poem's unique qualities derived from Ariosto's sustained inspiration and technique and detached ironical attitude to his characters. *Orlando furioso* was the most perfect expression of the literary tendencies of the Italian Renaissance at this time, and it exercised enormous influence on later European Renaissance literature. Ariosto also composed comedies that, by introducing imitation of Latin comedy, marked the beginning of Renaissance drama in the vernacular.

There were also attempts to renew the epic by submitting the tradition of chivalry to Aristotle's rules of composition. Gian Giorgio Trissino, a theorist on language, wrote according to the strictest Aristotelian rules, while Luigi Alamanni tried to focus the narrative on a single character in *Girone il cortese* (1548; "Girone the Courteous") and *Avarchide* (1570), an imitation of the *Iliad* of Homer. Giambattista Giraldi, while more famous as a tragic playwright, was a literary theorist who tried to apply his theories to his own poem *Ercole* (1557; "Hercules").

Two burlesque medley forms of verse were invented during the century. *Fidenziana* poetry derives its name from a work by Camillo Scroffa, a poet who wrote in a combination of Latin words and Italian form and syntax. Macaronic poetry, on the other hand, is a term given to verse consisting of Italian words used according to Latin form and syntax. Teofilo Folengo, a Benedictine monk, was the best representative of macaronic literature, and his masterpiece was a poem in 20 books called *Baldus* (1517). A tendency to parody, ridiculing the excesses of humanist literature, was present in both the *fidenziana* and macaronic verse.

Fidenziana and macaronic poetry

Torquato Tasso was the last great poet of the Italian Renaissance and one of the greatest of Italian literature. In his epic *Gerusalemme liberata* (1581; *Jerusalem Delivered*) he summed up a literary tradition typical of the Renaissance: the classical epic renewed according to the spiritual interests of his own time. Whereas most of Tasso's work shows a conflict between a desire to express himself according to classical ideals and a tendency to moralize, some of his works reflected his spontaneous inspiration. *L'Aminta* (1573), a joyous and uninhibited drama, was the best example of Tasso's youthful poetry and belonged to the new literary genre of pastoral (dealing with idealized rural life). *Gerusalemme liberata*, however, was the result of a balance in the poet's conflicting aspirations: a Christian subject dealt with in a classical way. In the subsequent *Gerusalemme conquistata* (1593; "Jerusalem Vanquished"), Tasso recast his poem according to strict Aristotelian rules and the ideals of the Roman Catholic Church's reaction against the Protestant Reformation, known as the Counter-Reformation. Tasso's conflict had ended in the victory of the moralistic principle: poetically the new poem was a failure. Tasso also wrote shorter verses throughout his life, including religious poems, and his prose works show a style no longer exclusively dominated by classical models.

Drama. Trissino's *Sofonisba* (written 1514–15) was the first tragedy of Italian vernacular literature; its structure derived from Greek models, but its poetic qualities were somewhat mediocre. Toward the middle of the 16th century Giambattista Giraldi (Cinzio) reacted against imitation of Greek drama by proposing the Roman tragedian Seneca as a new model, and in nine tragedies and tragicomedies—written between 1541 and 1549—he showed some independence from Aristotelian rules. He greatly influenced European drama, particularly the English of the Elizabethan period.

Italian comedies

Italian comedies of the century, inspired by Latin models, possessed greater artistic value than tragedies, and they reflected contemporary life more fully: they could be considered as the starting point for modern European drama. To the comedies of Ariosto and Machiavelli should be added a lively play, *La Calandria* (first performed 1513; *The Follies of Calandro*), by Cardinal Bernardo Dovizi da Bibbiena, and five equally amusing comedies written by Pietro Aretino. Giordano Bruno, a great Italian philosopher who wrote dialogues in Italian on his new cosmology and antihumanist ideas, also wrote a comedy, *Il candelaio* (1582; *The Candlemaker*).

Since the mid-20th century Angelo Beolco ("Il Ruzzante") has become generally recognized as the most powerful dramatist of the 16th century. His works, written in rustic Paduan dialect, treat the problems of the countryside with profound seriousness. Another dialect playwright of the same century, now also more widely appreciated, is the Venetian Andrea Calmo, who showed a nice gift for characterization in his comedies of complex amorous intrigue.

Narrative. The classicist trend established by Pietro Bembo also affected narrative literature, for which the obvious model was Boccaccio's *Decameron.* Originality and liveliness of expression were to be found in the 22 stories called *Le cene* (written after 1549; "The Suppers") of the Florentine apothecary Anton Francesco Grazzini. The worldly monk Agnolo Firenzuola produced several stories, including the fable *Asino d'oro* (1550), a free version of Apuleius' *Golden Ass.* The cleric and short-story writer Matteo Bandello started a new trend in 16th-century narrative with 214 stories that were rich in dramatic and romantic elements while not aiming at classical dignity. This trend was partly followed also by Giraldi in his collection of 112 stories called *Gli ecatommiti* (1565; "The Hundred Stories").

17TH-CENTURY LITERATURE

The 17th century in Italian literature is usually described as a period of "decadence" in which writers who were devoid of sentiment resorted to exaggeration and tried to cloak an utter poverty of matter beneath an exuberance of form. (In this period, it is said, freedom of thought was fettered by the Accademia della Crusca of Florence, whose aim it was to maintain the purity of the Tuscan tongue; by the Counter-Reformation; and by the political supremacy of Spain.) This style of writing was not, however, simply an Italian phenomenon. It was at this time that "Gongorism" (the ingenious, metaphorical style of the poet Luis de Góngora) flourished in Spain, and the witty, figurative verse of the Metaphysical poets was popular in England. Far from being exhausted, indeed, this was an extremely vital period, so much so that a new and more comprehensive understanding of the literature of the Italian Baroque has been formulated by scholars conversant with the changing attitude toward this phase of civilization in Germany, France, and England.

Poetry and prose. The popularity of satire was a reaction against prevailing conditions. Prominent in this genre was Salvator Rosa, who attacked in seven satires the vices and shortcomings of the age. Alessandro Tassoni acquired great fame with *La secchia rapita* (1622; *The Rape of the Bucket*), a mock-heroic poem that is both an epic and a personal satire. The greatest poet of the period was Tommaso Campanella, a Dominican friar, who is less well-known for his rough-hewn, philosophical verses than for the *Città del sole* (1602; *Campanella's City of the Sun*), a vision of political utopia, in which he advocated the uniting of humanity under a theocracy based on natural religion.

The influence of Marino

The principal representative of Italian writing during this period was Giambattista Marino, author of a large collection of lyric verse (*La lira* [1608–14; "The Lyre"] and *La sampogna* [1620; "The Syrinx"]), and a long mythological poem, *Adone* (1623). Marino derived inspiration from the poetry of the late 16th century, but his aim—typical of the age—was to excite wonder by novelty. His work is characterized by "conceits" of fantastic ingenuity, farfetched metaphor, sensuality, extreme facility, and a superb technical skill. His imitators were innumerable, and most 17th-century Italian poets were influenced by his work.

Gabriello Chiabrera, soberer in style than Marino, was successful in imitating the metres of classical poetry (especially of the Greek Pindar) and excelled in the composition of musical canzonets (short lyrical verses). Toward the end of the century a patriotic sonneteer, Vincenzo da Filicaia, and Alessandro Guidi, who wrote exalted odes, were hailed as major poets, though Guidi's verse is now seen as little more than rhetoric.

Among prose writers of the period, the satirist Traiano Boccalini stood out with *Ragguagli di Parnasso* (1612–13; *Advertisements from Parnassus*) in the fight against Spanish domination. A history of the Council of Trent (which defined Catholic doctrines in reaction to the Reformation) was written by Paolo Sarpi, an advocate of the liberty of the Venetian state against papal interference, and a history of the rising of the Low Countries against Spain was written by Guido Bentivoglio. The Venetian novels of Girolamo Brusoni are still of interest, as are the travels of Pietro della Valle and the tales of Giambattista Basile. All the restless energy of this period reached its climax in the work of Galileo, a scientist who laid the foundations of mathematical philosophy and earned a prominent place in the history of Italian literature through the vigour and clarity of his prose.

The music drama; Accademia dell'Arcadia. With the rise of the music drama and the opera, Italian authors worked to an increasing extent with the lyric stage. Librettos written by poets such as Ottavio Rinuccini were planned with dramatic and musical artistry. During the 17th century a popular spirit entered the opera houses: intermezzi (short dramatic or musical light entertainments) were required between the acts, a practice that undermined the dramatic unity of the performance as a whole, and toward the end of the century every vestige of theatrical propriety was abandoned. The spread of Marino's influence was felt by many to be an abuse. In 1690 the Accademia dell'Arcadia was founded in Rome for the express purpose of eradicating "bad taste." The purpose of Arcadia was in tune with a genuinely felt need. Many of its members were rationalist followers of Descartes with severe classical sympathies, but their reaction consisted mainly in imitating the simplicity of the nymphs and shepherds who were supposed to have lived in the golden age, and thus a new artifice replaced an old one. A typical exponent of the Arcadian lyric was Pietro Metastasio, the 18th-century reformer of the operatic libretto.

18TH-CENTURY DEVELOPMENTS

Reform of the tragic theatre. In 1713 Francesco Scipione Maffei, an antiquary of Verona, produced *Merope*—a tragedy that met with great success and pointed the way toward reform of the Italian tragic theatre. Between 1726 and 1747 Antonio Conti—an admirer of Shakespeare—wrote four Roman tragedies in blank verse. It was not until 1782 and the writing of *Saul,* however, that an important Italian tragedian finally emerged in the person of Vittorio Alfieri. In strong contrast with Pietro Metastasio's and Paolo Rolli's *melodrammi*—librettos set to music or sometimes performed as plays in their own right—Alfieri's tragedies are harsh, bitter, and unmelodious. He chose classical and biblical themes, and through his hatred of tyranny and love of liberty he aspired to move his audience with magnanimous sentiments and patriotic fervour. Alfieri's influence in the Romantic period and the Risorgimento was immense, and, like Carlo Goldoni, he wrote an important autobiography, which gives a revealing account of his struggles to provide Italy with a corpus of drama comparable with that of the other European nations.

Goldoni's reform of the comedy. Metastasio's reform of the operatic libretto was paralleled in the mid-18th century by Goldoni's reform of comedy. Throughout the 17th century the commedia dell'arte—a colourful pantomime of improvisation, singing, mime, and acrobatics, often performed by actors of great virtuosity—had gradually replaced regular comedy, but by the early 18th century it had degenerated into mere buffoonery and obscenity with fixed characters and mannerisms. The dialogue was mostly improvised, and the plot—a complicated series of stage directions, known as the scenario—dealt mainly with forced marriages, star-crossed lovers, and the intrigues of servants and masters. Goldoni succeeded in replacing this traditional type of theatre with written works whose wit and vigour are especially evident when the Venetian scene is portrayed in a refined form of the local dialect. Perhaps because of his prolific output his work has sometimes been thought of as lacking in depth. His social observation is acute, however, and his characters are beautifully drawn. *La locandiera* (1753; "The Innkeeper"; Eng. trans., *Mirandolina*), with its heroine Mirandolina, a feminist before feminism's time, has things to say about class and the position of women that can still be appreciated today. Goldoni's rival and bitter controversialist, fellow Venetian Carlo Gozzi, also wrote comedy and satirical verse.

The world of learning. Giambattista Vico, Ludovico Antonio Muratori, Apostolo Zeno, and Scipione Maffei were writers who reflected the awakening of historical consciousness in Italy. Muratori collected the primary sources for the study of the Italian Middle Ages; Vico, in *Scienza nuova* (1725–44; *The New Science*), investigated the laws governing the progress of the human race and from the psychological study of man endeavoured to infer the laws by which civilizations rise, flourish, and fall. Giovanni Maria Mazzuchelli and Gerolamo Tiraboschi devoted themselves to literary history. Literary criticism also attracted attention; Gian Vincenzo Gravina, Vico, Maffei, Muratori, and several others, while advocating the imitation of the classics, realized that it should be cautious and thus anticipated critical standpoints that were later to come into favour.

The Enlightenment. With the end of Spanish domination and the spread of the ideas of the Enlightenment from France, reforms were gradually introduced in various parts of Italy. The new spirit of the times led men—mainly of the upper middle class—to enquire into the mechanics of economic and social laws. The ideas and aspirations of the Enlightenment as a whole were effectively voiced in such organs of the new journalism as Pietro Verri's periodical *Il Caffè* (1764–66; "The Coffeehouse"). A notable contributor to *Il Caffè* was the philosopher and economist Cesare Beccaria, who in his pioneering book *Dei delitti e delle pene* (1764; *On Crimes and Punishments*) made an eloquent plea for the abolition of torture and the death penalty.

Parini and the literary revival

More than anyone else, Giuseppe Parini seems to embody the literary revival of the 18th century. In *Il giorno* (published in four parts, 1763–1801; "The Day"), a long social satire of the rights of blood, he described a day in the life of a young Milanese patrician and revealed with masterly irony the irresponsibility and futility of a whole way of life. His *Odi* (1795; "Odes"), which are imbued with the same spirit of moral and social reform, are among the classics of Italian poetry.

The satire in the *Sermoni* (1763; "Sermons") of Gasparo Gozzi (elder brother of Carlo) is less pungent, though directed at similar ends, and in his two periodicals—*La Gazzetta veneta* and *L'Osservatore*—he presented a lively chronicle of Venetian life and indicated a practical moral with much good sense. Giuseppe Baretti—an extremely controversial figure who published a critical journal called *La Frusta letteraria* ("The Literary Whip"), in which he castigated "bad authors"—had learned much through a lengthy sojourn in England, where his friendship with Samuel Johnson helped to give independence and vigour, if not always accuracy, to his judgments. The *Viaggi di Enrico Wanton* (1749–64), a philosophical novel in the form of an imaginary voyage by the Venetian Zaccaria Seriman, was the most all-embracing satire of the time.

LITERARY TRENDS OF THE 19TH CENTURY

The 19th century was a period of political ferment in Italy, and many outstanding writers were involved in public affairs. Much of the literature written with a political aim, even when not of intrinsic value, became part of Italy's national heritage and inspired not only those for whom it was written but all who valued freedom.

Romanticism. Foremost among writers in early struggles for his country's unity and freedom from foreign domination was Ugo Foscolo, who reconciled passionate feeling with a formal perfection inspired by classical models. His *Ultime lettere di Jacopo Ortis* (1802; "Jacopo Ortis' Last Letters") was an epistolary story of a young man forced to suicide by frustrated love for both a woman and his fatherland. It was extremely moving and popular, as was a poem, "Dei sepolcri" (1807; "On Sepulchres"), where, in fewer than 300 lines, he wrote lyrically on the theme of inspiration to be found at the tombs of the great, exhorting Italians to be worthy of their heritage. This poem influenced the Italian Risorgimento, or national revival, and a passage in which Florence was praised because it kept in Sta. Croce the ashes of Michelangelo, Machiavelli, and Galileo is still very popular in Italy. Two odes celebrating the divine quality of beauty, 12 sonnets ranking with the best of Petrarch's and Tasso's, and an unfinished poem, "Le grazie" ("The Graces"), also testified to Foscolo's outstanding poetic merit. As an exile in England from 1816 until his death in 1827, he wrote remarkable critical essays on Italian literature for English readers.

In Foscolo patriotism and classicism united to form almost one passion, but Vincenzo Monti was outstanding for mobility of feeling. He saw danger to his country in the French Revolution and wrote "Il pellegrino apostolico" (1782; "The Apostolic Pilgrim") and *La bassvilliana* (1793; *The Penance of Hugo*); Napoleon's victories aroused his praise in "Prometeo" (*c.* 1805; "Prometheus"), "Il bardo della selva nera" (1806; "The Bard of the Dark Wood"), and "La spada di Federico II" (1806; "The Sword of Frederick II"); in "Il fanatismo" and "La superstizione" (1797) he attacked the papacy; later he extolled the Austrians. Thus every great event made him change his mind, through lack of political conviction, yet he achieved greatness in *La bellezza dell'universo* (1781; "The Beauty of the Universe"), the lyrics inspired by domestic affections, and in a translation of the *Iliad,* a masterpiece of Neoclassical beauty.

Classicism, purism, and the reaction of Romanticism

Melchiorre Cesarotti occupied a prominent position in the world of learning at the end of the 18th century, and his translations of James Macpherson's Ossian poetry, *Poesie di Ossian* (1763–72), influenced Foscolo, Leopardi, and others by their mysterious and gloomy fantasy, so alien from classical inspiration; *Saggio sulla filosofia delle lingue* (1785; "Essay on the Philosophy of Languages") was an important essay in the dispute on the Italian language. The trend was toward pedantic classicism as a reaction against an excessive Gallicism favoured by some 18th-century writers. Among the purists was Antonio Cesari, who brought out a new enlarged edition of the *Vocabolario della Crusca* (the first Italian dictionary, published by the Accademia della Crusca in 1612). He wrote *Sopra lo stato presente della lingua italiana* (1810; "On the Present State of the Italian Language") and endeavoured to establish the supremacy of Tuscan and of Dante, Petrarch, and Boccaccio as models. But a Lombard school opposed this Tuscan supremacy. Monti, its leader, issued *Proposta di alcune correzioni ed aggiunte al vocabolario della Crusca* (1817–26; "Proposal for Some Corrections and Additions to the Crusca Dictionary"), which attacked the Tuscanism of the Crusca. By contrast, the patriot Pietro Giordani—for a time a journalistic colleague of Monti—was a great exponent of *purismo.* His views did not stem from literary pedantry, however, but from a concern that all social groups throughout Italy should have a common means of communication. In this respect he was linguistically opposed to the poet Carlo Porta, who lampooned the aristocracy and clergy and expressed sympathy with the humble and wretched in a lively Milanese dialect. All Italy took part in the disputes about language, literature, and politics.

An artificial form of classicism was associated with the Napoleonic domination of Italy, so that when Napoleon fell, forces antagonistic to classicism arose. Literary Romanticism had already won favour with the French, who erroneously thought themselves akin to German Romantics. Between 1816 and 1818 a battle was fought for Romanticism, particularly in Milan, where a Romantic periodical, *Il Conciliatore* (1818–19; "The Peacemaker"), was published. Giovanni Berchet (patriotic poet and author of *La lettera semiseria di Grisostomo,* a manifesto of Romanticism published in 1816), Silvio Pellico, Ludovico di Breme, Giovita Scalvini, and Ermes Visconti were among its contributors. Their efforts were silenced in 1820 when several were arrested by the Austrian police because of their liberal opinions; among them was Pellico, who later wrote a famous account of his experiences, *Le mie prigioni* (1832; *My Prisons*).

The publication of *Il Conciliatore*

Alessandro Manzoni (grandson of Cesare Beccaria) was the chief exponent of Italian Romanticism, but perhaps an even higher claim to fame was his contribution to the resolution of the language problem. In 1821 he started working on a panoramic novel about the lives of simple people placed against a background of major historical events, and, in order that this should be accessible to a wide readership, he decided to write it in an idiom as close as possible to modern educated Florentine speech. This was a formidable enterprise for someone whose first languages were French and Milanese dialect—and to whom spoken Florentine was virtually a foreign tongue—and for the first draft (completed in 1823) he had to resort to Francesco Cherubini's Italian-Milanese dictionary. The second draft was published in 1825–27 under the title *I promessi sposi* (*The Betrothed*); and the final definitive edition came out in 1840–42 after a long, painstaking process of revision aimed at making the text conform more closely with colloquial Florentine usage. The result of this effort was clear, expressive prose—neither pretentious nor provincial—and the way in which the novel caught the public's imagination attested to Manzoni's success in addressing the sort of people to whom conventional literary Italian was almost as remote as Latin. Ironically, Manzoni the innovator became, in his turn, the model for a new kind of purism, with "Manzonians" composing works in an affected Tuscan, and it required authors with fresh ideas—not poor imitators—to continue the task of disencumbering and modernizing written Italian.

Manzoni's genius as a poet showed in the odes "Il cinque maggio" (1821; "The Fifth of May"; Eng. trans., "The Napoleonic Ode"), written on the death of Napoleon, and "Marzo 1821," and in passages of *Inni sacri* (1812–22; *Sacred Hymns*), five poems in celebration of church festivals, describing human affections. His tragedies, *Il conte di Carmagnola* (performed 1820; "The Count of Carmagnola") and *Adelchi* (1822), marked a victory of Romanticism over classicism; they contained passages of great lyrical beauty but lacked strong dramatic power.

The foremost Italian poet of the age was Giacomo Leopardi, an outstanding scholar and thinker whose philological works together with his philosophical writings, *Operette morali,* would alone place him among the great writers of the 19th century. Embittered by solitude, sickness, and near penury, from the age of 20 he realized the vanity of hope. Though he developed a doctrine of universal pessimism, seeing life as evil and death as the only comfort, the poetry based on these bitter, despairing premises was far from depressing. Most of Leopardi's poems were contained in one book, *I canti* ("Songs"; Eng. trans., *The Poems of Leopardi*), first published in 1831. Some were patriotic and were once very popular; but the best came from deeper lyrical inspiration. Among them were a meditation on infinity; "A Silvia," on the memory of a girl who died when he was 20; "Le ricordanze," an evocation of his childhood; "Il passero solitario," comparing the lonely poet and a sparrow that sings by itself; and "La quiete dopo la tempesta" and "Il sabato del villaggio," two pictures of village life. They balance depth of meaning and formal beauty, simplicity of diction, intensity, and verbal music.

Leopardi's pessimism

The Risorgimento and after. Circumstances made it inevitable that Italian Romanticism should become heavily involved with the patriotic myths of the Risorgimento; yet, while this served a useful civic purpose at the time, it did not encourage literature of consistent artistic merit or enduring readability. Of the writings produced by figures associated in some way with Italy's struggle for nationhood, it tends to be the less typical ones that attract attention today: the dialect poetry of Giuseppe Gioacchino Belli describing the life of contemporary papal Rome; verses by Giuseppe Giusti satirizing petty tyrants, political turncoats, and coarse parvenus; or the works of the republican Roman Catholic from Dalmatia, Niccolò Tommaseo. The undoubted masterpiece of Risorgimento narrative literature is Ippolito Nievo's *Confessioni di un italiano* (published posthumously in 1867; "Confessions of an Italian"; Eng. trans., *The Castle of Fratta*), which marks Nievo as the most important novelist to emerge in the interval between Manzoni and Verga. Giuseppe Mazzini's letters can still be studied with profit, as can the memoirs of Luigi Settembrini (*Ricordanze della mia vita* [1879–80; "Recollections of My Life"]) and Massimo D'Azeglio (*I miei ricordi* [1868; *Things I Remember*]). D'Azeglio's historical novels or those of Francesco Guerrazzi now have a rather limited interest; and Mazzini's didactic writings—of great merit in their good intentions—are generally regarded as unduly oratorical. Giovanni Prati and Aleardo Aleardi, protagonists of the "Second Romanticism," wrote poetry of a sentimentality that helped to provoke a variety of reactive movements, including *scapigliatura* and *verismo.*

Giosuè Carducci was an outstanding figure whose enthusiastic support for the national cause during the struggle of 1859–61 was changed to disillusionment by the difficulties in which the new kingdom was involved. The bitterness of some of his poetry revealed frustration and rebelliousness. *Rime nuove* (*The New Lyrics*) and *Odi barbare* (*The Barbarian Odes*), both of which appeared in the 1880s, contained the best of his poetry: memories of childhood, evocations of landscape, laments for domestic sorrows, an inspired representation of historical events, an ambitious effort to resuscitate the glory of Roman history, and an anachronistic but sincere cult of pagan civilization. He tried to adapt Latin prosody to Italian verse, which sometimes produced good poems, but his opposition to Romanticism and his rhetorical tirades provoked a strong reaction, and his metrical reform was short-lived. He was also a scholarly historian of literature, and his literary essays had permanent value, although philosophical criticism such as that of De Sanctis was uncongenial to him. Both his poetry and his criticism were cited when he was awarded the Nobel Prize for Literature for 1906.

The works of Carducci

A figure connected politically with the Risorgimento but remembered chiefly for his critical writings was Francesco De Sanctis, whose most important works consisted of various critical essays and *Storia della letteratura italiana* (1870–71; *History of Italian Literature*). His main tenet was that literature was to be judged not on its intellectual or moralistic content so much as by the spirit of its "form," and the role of the critic was to discover how this form had been unconsciously and spontaneously conceived by studying its creator's temperament and background and the age in which he lived. De Sanctis was not properly appreciated in his day but came into his own at the turn of the century when Benedetto Croce rescued his works from oblivion.

While Carducci was still alive, Giovanni Pascoli acquired a reputation and succeeded him in the chair of Italian literature at the University of Bologna. His art was often impressionistic and fragmentary, his language occasionally laborious, but his lyricism, at first timid in inspiration in *Myricae* (1891; "Tamarisks"), rose to fuller tones when he attempted the loftier themes of antiquity: Roman heritage and greater Italy. His original vein still found expression in *Canti di Castelvecchio* (1903; "Songs of Castelvecchio") and in the classicism of *Poemi conviviali* (1904; "Convivial Poems"). Later he produced—both in humanistic Latin and in self-consciously elaborate Italian—heroic hymns in honour of two sacred cities, Rome and Turin.

The veristi and other narrative writers. The patriotic niceties and sentimental Romanticism of much Risorgimento writing inevitably provoked a reaction. The first serious opposition came from the *scapigliati* ("libertines" or "bohemians"), adherents of an antibourgeois literary and artistic movement that flourished in Milan and Turin during the last four decades of the 19th century and whose declared aim was to link up with the most advanced Romantic currents from abroad. Unfortunately the movement—perhaps by its very nature—lacked intellectual cohesion and tended to cultivate the eccentric as an end in itself. The *scapigliati,* however, made a useful contribution in social criticism and in their informal linguistic approach. Among the foremost *scapigliati* were Giuseppe Rovani, whose monumental novel about Milanese life, *I cento anni* (*The Hundred Years*), was issued in installments (1856–58 and 1864–65); Emilio Praga, a poet tormented by contradictions; and Arrigo Boito, poet, musician, and librettist for Giuseppe Verdi's *Falstaff* and *Otello.*

Scapigliati

A more lasting and fruitful successor to conventional Italian Romanticism was *verismo* ("realism"; first theoretically expounded by Luigi Capuana in 1872), a movement initially inspired by the French Naturalist writers and influenced by positivist and determinist ideas. The *veristi* were not concerned with sermons or noble sentiments but with observable phenomena. When they dealt with the Risorgimento, they showed it warts and all. The greatest of *verismo* narrators was without a doubt Giovanni Verga, who explained in a preamble to a short story, "L'amante di Gramigna" (1880; Eng. trans., "Gramigna's Lover"), that in a perfect novel the sincerity of its reality would be so evident that the hand of the artist would be absolutely invisible and the work of art would seem to have matured spontaneously without any point of contact with its author. At times Verga almost seems to have achieved this unattainable goal, and in his two great narrative works dealing with the victims of social and economic change, *I malavoglia* (1881; "The Unwilling"; Eng. trans., *The House by the Medlar Tree*) and *Mastro-don Gesualdo* (1889), the reader often has the sensation of being put down in an unfamiliar milieu and—as would happen in real life—left to pick up the threads from gossip and chance remarks. Another *verista,* Federico De Roberto, in his novel *I vicerè* (1894; *The Viceroys*), has given a cynical and wryly funny account of an aristocratic Sicilian family that adapted all too well to change. Luigi Capuana, the founder of *verismo* and most rigorous adherent to its impersonal method of narration, is known principally for his dramatic psychological study, *Il marchese di Roccaverdina* (1901; "The Marquis of Roccaverdina").

Verismo and *veristi*

In their search for documentary exactitude the *veristi* paid close attention to background. For Verga, De Roberto, and Capuana, this was Sicily. Matilde Serao, on the other hand, has given a detailed and colourful reportage of the Neapolitan scene, while Renato Fucini conveyed the atmosphere of traditional Tuscany. Emilio De Marchi, another writer in the realist mold, has Milan for his setting and in *Demetrio Pianelli* (1890) has painted a candid but essentially kindly portrait of the new Milanese petite bourgeoisie. Antonio Fogazzaro was akin to the *veristi* in his powers of observation and in his descriptions of minor characters; but he was strongly influenced by Manzoni, and his best narrative work, *Piccolo mondo antico* (1895; *The Little World of the Past*), is a nostalgic look back to a supposedly less individualistic age when inner tranquillity was seemingly achieved by devotion to a shared ideal. The *veristi* had a leavening effect on Italian literature generally, and their influence can be discerned, among others, in the early novels of the Sardinian Grazia Deledda (awarded the Nobel Prize for Literature for 1926) and in the distinguished narrative works of the Sienese writer Federigo Tozzi, including *Con gli occhi chiusi* (1919; "With Closed Eyes") and *Tre croci* (1920; *Three Crosses*).

THE 20TH CENTURY

Gabriele D'Annunzio's nationalism. After unification the new Italy was preoccupied with practical problems, and by the early 20th century a great deal of reasonably successful effort had been directed toward raising living standards, promoting social harmony, and healing the split between church and state. It was in this prosaic and pragmatic atmosphere that the middle classes—bored with the unheroic and positivist spirit of former decades—began to feel the need for a new myth. Thus it is easy to understand how imaginations across the political spectrum came to be fired by the extravagant personality of Gabriele D'Annunzio—man of action, nationalist, literary virtuoso, and (not least) exhibitionist—whose life and art seemed to be a blend of Jacob Burckhardt's "complete man" and the superman of Friedrich Nietzsche. At a distance from those times, it should be possible to evaluate D'Annunzio more clearly. There is, however, no critical consensus about his writings, although he is generally praised for his autobiographical novel, *Il piacere* (1889; *The Child of Pleasure*), for his mature poetry, and for his late memoirs.

Benedetto Croce's criticism. Although D'Annunzio's fame was worldwide, the function of modernizing intellectual life fell mainly to Benedetto Croce in almost 70 books and in the bimonthly review *La Critica* (1903–44). Perhaps his most influential work was literary criticism, which he expounded and continually revised in articles and books spanning nearly half a century.

Croce's beliefs implied condemnation of Fascism's ideology, but he was not seriously molested by the Fascist regime, and through the darkest days *La Critica* remained a source of encouragement to at least a restricted circle of freedom-loving intellectuals. Unfortunately, his highly systematized approach to criticism led to a certain rigidity and a refusal to recognize the merits of some obviously important writers, and this was undoubtedly one reason why after World War II his authority waned. His monumental corpus of philosophical, critical, and historical works of great scholarship, humour, and common sense remains, however, the greatest single intellectual feat in the history of modern Italian culture.

Literary trends before World War I. While Croce was starting his arduous task, literary life revolved mainly around reviews such as *Leonardo* (1903), *Hermes* (1904), *La Voce* (1908), and *Lacerba* (1913), founded and edited by relatively small groups. The two main literary trends were: *crepuscolarismo,* which favoured a colloquial style to express memories of sweet things past, as in the work of Guido Gozzano and Sergio Corazzini; and *futurismo,* loathing of traditional art and demanding complete freedom of expression, whose leader was Filippo Tommaso Marinetti, editor of *Poesia,* a fashionable cosmopolitan review. Both *crepuscolari* and *futuristi* were part of a complex European tradition of disillusionment and revolt, the former inheriting the sophisticated pessimism of French and Flemish "decadents," the latter taking part in an episode in the history of western European avant-garde developed from the French poets Stéphane Mallarmé and Guillaume Apollinaire to the Cubist, Surrealist, and Dada movements. Both shared a feeling of revulsion against D'Annunzian flamboyance and rhetoric, from which they attempted to free themselves and, paradoxically, from which they derived elements of their style (the "crepuscular" mood of D'Annunzio's *Poema paradisiaco* [1893]

Crepuscolarismo and *futurismo*

and most Futuristic "new theories"—identification of art with action, heroism, and speed, free use of words—being implied in his *Laus Vitae* [1903; "In Praise of Life"]).

The "return to order." The end of World War I saw a longing for the revival of tradition, summed up in the aims of the review *La Ronda,* founded in 1919 by Vincenzo Cardarelli and others, which advocated a return to classical stylistic values. This led to an excessive cult of form in the narrow sense—as exemplified by the elegant but somewhat bloodless essays (*elzeviri*) published in Italian newspapers on page three—and obviously fitted in with the stifling of free expression under Fascism. The sterility of this period, however, should not be exaggerated. The 20 years of Fascist rule were hardly conducive to creativity, but in the dark picture there were a few glimmers of light. With 1923 came the publication of Italo Svevo's *Coscienza di Zeno* (*The Confessions of Zeno*), a gem of psychological observation and Jewish humour, which a few years later was internationally "discovered" through the mediation of James Joyce. The surreal writings of Massimo Bontempelli (*Il figlio di due madri* [1929; "The Son of Two Mothers"]) and of Dino Buzzati (*Il deserto dei Tartari* [1940; *The Tartar Steppe*]) were perhaps in part an escape from the prevailing political climate, but they stand up artistically nonetheless. Riccardo Bacchelli, with *Il diavolo a Pontelungo* (1927; *The Devil at the Long Bridge*) and *Il mulino del Po* (1938–40; *The Mill on the Po*), produced historical narrative writing of lasting quality. Aldo Palazzeschi, in *Stampe dell'Ottocento* (1932; "Nineteenth Century Engravings") and *Sorelle Materassi* (1934; *The Sisters Materassi*), reached the height of his storytelling powers. Meanwhile, the Florentine literary reviews *Solaria, Frontespizio,* and *Letteratura,* while having to tread carefully with the authorities, provided an outlet for new talent. Carlo Emilio Gadda had his first narrative work (*La Madonna dei filosofi* [1931; "The Philosophers' Madonna"]) published in *Solaria,* while the first part of his masterpiece, *La cognizione del dolore* (*Acquainted with Grief*), was serialized between 1938 and 1941 in *Letteratura.* Novelists such as Alberto Moravia, Corrado Alvaro (*Gente in Aspromonte* [1930; *Revolt in Aspromonte*]), and Carlo Bernari had to use circumspection in stating their views but were not completely silenced. Ignazio Silone, having chosen exile, could speak openly in *Fontamara* (1930). Antonio Gramsci, an unwilling "guest" of the regime, gave testimony to the triumph of spirit over oppression in *Lettere dal carcere* (1947; *Letters from Prison*).

Sterility and creativity under Fascism

Luigi Pirandello. Drama, which a few playwrights and producers were trying to extricate from old-fashioned realistic formulas and more recent superhuman theories, was increasingly dominated by Luigi Pirandello. His own experience of the "unreal," through his calamitous family life and his wife's insanity, enabled him to see the limitations of realism. From initial short-story writing, in which he explored the incoherence of personality, the lack of communication between individuals, the uncertain boundaries between sanity and insanity or reality and appearance, and the relativity of truth, he turned to drama as a better means of expressing life's absurdity and the ambiguous relationship between fact and fiction.

Pirandello's play within a play

To multiply the fragmentation of the levels of reality, Pirandello tried to destroy conventional dramatic structures and to adopt new ones: a play within a play in *Sei personaggi in cerca d'autore* (1921; *Six Characters in Search of an Author*) and a scripted improvisation in *Questa sera si recita a soggetto* (1930; *Tonight We Improvise*). This was a way of transferring the dissociation of reality from the plane of content to that of form, thereby achieving an almost perfect unity between ideas and dramatic structure. Pirandello's plays, including perhaps his best, *Enrico IV* (1922; *Henry IV*), often contain logical arguments: several critics, including Croce, were misled into thinking that he intended to express in this way a coherent philosophy, whereas he used logic as a dramatic symbol. Pirandello was awarded the 1934 Nobel Prize for Literature.

The Hermetic movement. Poetry in the Fascist period underwent a process of involution, partly influenced by French Symbolism, with its faith in the mystical power of words, and partly under the stress of changed political conditions after World War I, during which literature had declined. Many poets of the wartime generation, weary of tradition and rhetoric, had been seeking new expression: some, like the *futuristi,* had tried to work rhetoric out of their system by letting it run amok; others, such as Camillo Sbarbaro (*Pianissimo* [1914], *Trucioli* [1920; "Shavings"]), cultivated a style purified of unessential elements. Out of those efforts grew a poetry combining the acoustic potentialities of words with emotional restraint and consisting mainly of fragmentary utterances in which words were enhanced by contextual isolation and disruption of syntactic and semantic links. The resultant obscurity compensated poets for loss of influence in a society subservient to dictatorship by turning them into an elite and allowed some, notably Eugenio Montale (who won the Nobel Prize for Literature for 1975), to express their pessimism covertly. The name of this movement, *ermetismo* ("Hermeticism"), hinted at both its aristocratic ambitions and its esoteric theory and practice. Its leader, Giuseppe Ungaretti, tried to charge each word of his early poems with such intensity of meaning that concern with technical problems often overshadowed emotion, thus producing supremely stylized forms. Thus, what in the 1920s had appeared revolutionary proved later to be only another facet of the formalistic tradition. Against this background of refinement, obscurity, and unreality, only the simple and moving poems of Umberto Saba preserved an immediate appeal.

Experiments in obscurity

Social commitment and the new realism. During World War II the walls of the hermetic ivory tower began to crumble. Ungaretti's style became so articulate as to be almost unrecognizable. Salvatore Quasimodo adopted a new engagé, or committed, style, which won critical admiration, including the 1959 Nobel Prize for Literature, and others followed suit in a drift toward social realism.

Social realism

This development had been foreshadowed by some writers under Fascism. In 1929 Alberto Moravia had written a scathing indictment of middle-class moral indifference, *Gli indifferenti* (1929; *Time of Indifference*). Carlo Bernari wrote a novel about the working classes, *Tre operai* (1934; "Three Workmen"); Cesare Pavese produced *Paesi tuoi* (1941; "Your Lands"; Eng. trans., *The Harvesters*); and Elio Vittorini wrote *Conversazione in Sicilia* (1941; *Conversation in Sicily*); all definitely promised a new literary development. From these and from aspects of American literature (William Faulkner, Erskine Caldwell, John Steinbeck, John Dos Passos, and Ernest Hemingway, translated mainly by Elio Vittorini and Pavese) postwar writing took its cue. Certain English literature, the homegrown *veristi,* and the ideas of Marxism were also an influence on postwar authors, to whom in varying degrees the rather imprecise label of Neorealism was attached. It was a stimulating time in which to write, with a wealth of unused material at hand. There were the social and economic problems of the south, described by Carlo Levi in his poetic portrait of Lucania, *Cristo si è fermato a Eboli* (1945; *Christ Stopped at Eboli*), and by Rocco Scotellaro (*Contadini del sud* [1954; "Peasants of the South"]) and Francesco Jovine (*Le terre del Sacramento* [1950; "The Lands of the Sacrament"; Eng. trans., *The Estate in Abruzzi*]). Vivid pictures of the Florentine working classes were painted by Vasco Pratolini (*Il quartiere* [1945; "The District"; Eng. trans., *The Naked Streets*] and *Metello* [1955]) and of the Roman subproletariat by Pier Paolo Pasolini (*Ragazzi di vita* [1955; *The Ragazzi*] and *Una vita violenta* [1959; *A Violent Life*]). There were memories of the north's struggle against Fascist and Nazi domination from Vittorini and from Beppe Fenoglio (*I ventitrè giorni della città di Alba* [1952; "The 23 Days of the City of Alba"]). There were sad tales of lost war by Giuseppe Berto (*Il cielo è rosso* [1947; *The Sky Is Red*] and *Guerra in camicia nera* [1955; "A Blackshirt's War"]) and by Mario Rigoni Stern (*Il sergente nella neve* [1952; "The Sergeant in the Snow"; Eng. trans. in *The Lost Legions; Three Italian War Novels*]). By contrast, there were humorous recollections of provincial life under Fascism—for example, Mario Tobino's *Bandiera nera* (1950; "Black Flag") and Goffredo Parise's *Prete bello* (1954; "The Handsome Priest"; Eng. trans., *The Priest Among the Pigeons*). In contrast to the more topical appeal of these writings the great virtue of

Pavese's narrative was the universality of its characters and themes. Among his finest works may be numbered *La casa in collina* (1949; *The House on the Hill*) and *La luna e i falò* (1950; *The Moon and the Bonfires*). Also of lasting relevance is Primo Levi's moving account of how human dignity survived the degradations of Auschwitz (*Se questo è un uomo* [1947; *If This Is a Man*]).

Other writings. Literary tastes gradually became less homogeneous. On the one hand, there was the rediscovery of the experimentalism of Carlo Emilio Gadda, whose best works had been written between 1938 and 1947. On the other, there was the runaway success of Giuseppe Tomasi di Lampedusa's old-fashioned historical novel *Il gattopardo* (1958; *The Leopard*), a soft-focused, flattering view of a family similar to the one described so pitilessly by Federico De Roberto in *I vicerè*. For this reason, it is easier to see Italian writing in terms of individual territory rather than general trends. Carlo Cassola's most memorable novels use the stillness of rural Tuscany as a background to the interior reality of its inhabitants, and in this his lineage can be traced to other Tuscan writers such as Romano Bilenchi (*La siccità* [1941; "The Drought"]) and Nicola Lisi (*Diario di un parroco di campagna* [1942; "Diary of a Country Priest"]) or in some respects back to Federigo Tozzi. Especially typical of Cassola's works are *Il taglio del bosco* (1953; *The Felling of the Forest*), *Un cuore arido* (1961; *An Arid Heart*), and *Un uomo solo* (1978; "A Man by Himself"). Giorgio Bassani's domain is the sadly nostalgic world of Ferrara in days gone by, with particular emphasis on its Jewish community (*Il giardino dei Finzi-Contini* [1962; *The Garden of the Finzi-Continis*]). Italo Calvino concentrated on fantastic tales (*Il visconte dimezzato* [1952; *The Cloven Viscount*], *Il barone rampante* [1957; *The Baron in the Trees*], and *Il cavaliere inesistente* [1959; *The Nonexistent Knight*]—all three are translated in *Our Ancestors: Three Novels* [1980])—and, later, on moralizing science fiction (*Le cosmicomiche* [1965; *Cosmicomics*] and *Ti con zero* [1968; *t zero*]). Paolo Volponi's province is the human consequences of Italy's rapid postwar industrialization (*Memoriale* [1962; *The Memorandum*], *La macchina mondiale* [1965; *The Worldwide Machine*], and *Corporale* [1974]). Leonardo Sciascia's sphere is his native Sicily, whose present and past he displays with concerned and scholarly insight, with two of his better-known books—in the format of thrillers—covering the sinister operations of the local Mafia (*Il giorno della civetta* [1963; *The Day of the Owl*] and *A ciascuno il suo* [1966; *To Each His Own*, also translated as *A Man's Blessing*]). After a Neorealistic phase, Giuseppe Berto plunged into the world of psychological introspection (*Il male oscuro* [1964; "The Dark Sickness," Eng. trans. *Incubus*] and *La cosa buffa* [1966; "The Funny Thing," Eng. trans. *Antonio in Love*]). Natalia Ginzburg's territory is the family, whether she reminisces about her own (*Lessico famigliare* [1963; *Family Sayings*, also trans. as *The Things We Used to Say*]), handles fictional characters (as in *Famiglia* [1977; *Family*]), or ventures into historical biography (*La famiglia Manzoni* [1983; *The Manzoni Family*]). Giovanni Arpino excelled at personal sympathies that cross cultural boundaries (*La suora giovane* [1959; *The Novice*] and *Il fratello italiano* [1980; "The Italian Brother"]). Fulvio Tomizza also tackled this theme in *L'amicizia* (1980; "The Friendship"). Meanwhile, Alberto Moravia and Mario Soldati defended their corners as never less than conspicuously competent writers. Moravia generally plowed a lone furrow. Of his mature writings, *Agostino* (1944), *Il conformista* (1951; *The Conformist*), and *La noia* (1960; *Boredom*, also translated as *Empty Canvas*) stand out as particular achievements. Soldati, in works such as *Le lettere da Capri* (1953; *The Capri Letters*) and *Le due città* (1964; "The Two Cities," Eng. trans. *The Malacca Cane*)—and in a later novel, *L'incendio* (1981; "The Fire"), which takes a quizzical look at the modern art business—showed himself to be a consistently skilled and entertaining narrator. There are many other accomplished authors who could be classified in this way, including Elsa Morante, who with *L'isola de Arturo* (1957; *Arturo's Island*) and *La storia* (1974; *History*) carved a unique niche for herself. Set in Rome during the years 1941–47, the combination of fact and allegory is a tour de force and one of the most remarkable narrative works to have come out of Italy since World War II.

Among the major novelists of the latter half of the 20th century was Umberto Eco, also a critic and semiotician, who wrote the internationally acclaimed *Il nome della rosa* (1981; *The Name of the Rose*), a murder mystery set in a 14th-century Italian monastery but also a questioning of "truth" from several perspectives. Calvino's later works, such as *Le città invisibili* (1972; *Invisible Cities*) and *Se un notte d'inverno un viaggiatore* (1979; *If on a Winter's Night a Traveler*), confirmed him as one of the most important Italian fiction writers of the century. Anna Maria Ortese wrote complex romantic tales that have been compared to those of Isak Dinesen. Poets of note included the lyrically autobiographical Attilio Bertolucci (father of the film director Bernardo Bertolucci); the psychoanalytically trained Cesare Viviani; the Symbolist Lorenzo Calogero; and the acclaimed experimentalist Andrea Zanzotto, whose volumes include the bilingual collection *Peasant's Wake for Fellini's Casanova and Other Poems* (1997). Among the dramatists of the late 20th century was the satirical and comedic Dario Fo, who won the 1997 Nobel Prize for Literature. One of Fo's later plays is *Il papa e la strega* (1989; *The Pope and the Witch*).

(G.A./S.Ra./G.P.Gi./D.M.Wh./F.Do./G.Car./Jo.M./Ed.)

BIBLIOGRAPHY. Literary histories include FRANCESCO DE SANCTIS, *History of Italian Literature*, trans. by JOAN REDFERN, 2 vol. (1930, trans. from Italian new ed. 1912); ROBERT ANDERSON HALL, *A Short History of Italian Literature* (1951); ERNEST HATCH WILKINS, *A History of Italian Literature*, rev. by THOMAS G. BERGIN (1974); and J.H. WHITFIELD, *A Short History of Italian Literature*, 2nd ed. (1980). PETER BONDANELLA and JULIA CONAWAY BONDANELLA (eds.), *Dictionary of Italian Literature*, rev., expanded ed. (1996), is a guide to authors, genres, schools, and periods of Italian literary history.

Studies that concentrate on specific periods or trends of Italian literary history include EZRA POUND, *The Spirit of Romance* (1910, reissued 1968); JOHN FRECCERO, *Dante: The Poetics of Conversion*, ed. by RACHEL JACOFF (1986); GIUSEPPE MAZZOTTA, *The World at Play in Boccaccio's Decameron* (1986); THOMAS M. GREENE, *The Light in Troy: Imitation and Discovery in Renaissance Poetry* (1982, reprinted 1986); DAVID MARSH, *The Quattrocento Dialogue: Classical Tradition and Humanist Innovation* (1980); BERNARD WEINBERG, *A History of Literary Criticism in the Italian Renaissance*, 2 vol. (1961, reprinted 1974); LOUISE GEORGE CLUBB, *Italian Drama in Shakespeare's Time* (1989); ALBERT RUSSELL ASCOLI, *Ariosto's Bitter Harmony: Crisis and Evasion in the Italian Renaissance* (1987); CONSTANCE JORDAN, *Renaissance Feminism: Literary Texts and Political Models* (1990); ANN ROSALIND JONES, *The Currency of Eros: Women's Love Lyric in Europe, 1540–1620* (1990); GLAUCO CAMBON, *Ugo Foscolo, Poet of Exile* (1980); DINO CARPANETTO and GIUSEPPE RICUPERATI, *Italy in the Age of Reason, 1685–1789*, trans. from Italian (1987); CAROLYN SPRINGER, *The Marble Wilderness: Ruins and Representation in Italian Romanticism, 1775–1850* (1987); REBECCA J. WEST, *Eugenio Montale, Poet on the Edge* (1981); SERGIO PACIFICI, *A Guide to Contemporary Italian Literature, from Futurism to Neorealism* (1962, reissued 1972), and *The Modern Italian Novel*, 3 vol. (1967–79); DONALD HEINEY, *Three Italian Novelists: Moravia, Pavese, Vittorini* (1968); JOHN GATT-RUTTER, *Writers and Politics in Modern Italy* (1978), a study of the relationship between politics and literature since World War II; MICHAEL CAESAR and PETER HAINSWORTH (eds.), *Writers & Society in Contemporary Italy: A Collection of Essays* (1984, reissued 1986); and MICHELE RICCIARDELLI, *Writings on Twentieth-Century Italian Literature, 1964–1984* (1992), offering useful profiles, reviews, and interviews; JOSEPH CARY, *Three Modern Italian Poets: Saba, Ungaretti, Montale*, 2nd ed. (1993); and THOMAS E. PETERSON, *The Rose in Contemporary Italian Poetry* (2000). Specific areas of extraliterary interest are treated in MILLICENT MARCUS, *Filmmaking by the Book: Italian Cinema and Literary Adaptation* (1993); SANTO L. ARICÒ (ed.), *Contemporary Women Writers in Italy* (1990); and LYNN M. GUNZBERG, *Strangers at Home: Jews in the Italian Literary Imagination* (1992).

Italy

With a shape that has been likened frequently to a high-heeled boot about to prod its triangular subject island of Sicily, the peninsular home of the European nation of Italy (officially Italian Republic; Italian: Italia, or Repubblica Italiana) juts deep into the Mediterranean Sea. Another important island, Sardinia, lies some 160 miles (260 kilometres) west of Italy's "shin." The magnificent mountain barrier of the Alps forms a northern boundary, which historically has hindered marauders less than might be supposed; these mountains separate Italy from France, Switzerland, Austria, and Slovenia and extend all the way down the Italian peninsula as a less elevated chain, the Apennines. Areas of plain, which are practically limited to the great northern triangle of the Po Valley, cover a mere 21 percent of the total national area of 116,000 square miles (301,000 square kilometres); 40 percent is hilly and 39 percent mountainous, providing variations to the generally temperate climate.

The mountainous landscape of Italy has long influenced political and economic developments on the peninsula by encouraging the creation of numerous independent states and by permitting in many regions only a meagre agriculture, providing grain sufficient only for a subsistence economy. Increased cultivation has caused deforestation. Since World War II, increasing numbers of Italians have abandoned the countryside for the rapidly industrializing cities, often creating severe dislocations in traditional ways of life.

The Italian economy, now ranked high in the world, blends areas as diverse as the "industrial triangle," formed by Milan, Turin, and Genoa, dating from about 1900, and the backward regions of the south and the islands, which are, however, being developed, mostly with aid from the state and the European Union.

Agriculture, which operates in often difficult natural and economic conditions, contributes about 4 percent of the gross domestic product (GDP), industry about 30 percent, and public and private services more than 50 percent. Sufficient wheat is grown for the population, and vegetables, fruit, grapes, and olives are cultivated in suitable districts. Cattle raising, however, is less advanced; meat and dairy produce is imported.

Italian industry includes every type of production. Although mineral resources are scarce, imported raw materials since World War II have boosted the production of iron and steel, other metallurgy, and construction. The chemical industry also flourishes, and textiles constitute one of Italy's largest industries. Services, particularly tourism, are very important, and efforts have been made to provide comprehensive networks of *autostrade* (express highways).

The peninsula has a proud tradition dating from antiquity. From its unification in the second half of the 19th century until 1946, Italy was a monarchy. Then it became a parliamentary republic, operating under the constitution of 1948. The republic is subdivided into *regioni, province,* and *comuni* ("communes"); these local bodies enjoy a certain autonomy, especially the regions, which differ widely in economic development. A similar diversity characterizes political life. From the end of World War II to the early 1990s, Italy had a multiparty system dominated by two large parties—the Christian Democratic Party (Partito della Democrazia Cristiana; DC) and the Italian Communist Party (Partito Comunista Italiano; PCI)—and a number of small but influential parties. The DC was the dominant governing party, in various alliances with the smaller parties of the centre and left. The Italian party system underwent a radical transformation in the early 1990s as a result of both international and national events. In 1991 the Communist Party became the Democratic Party of the Left (Partito Democratico della Sinistra; PDS). The DC disappeared altogether. One of its successors was the much weaker Italian Popular Party (Partito Popolare Italiano; PPI). The main result of these changes was the collapse of the political centre and a right-left polarization of the party spectrum.

Workers' unions have been an important part of national life. They are grounded in various confederations, principally the Italian General Confederation of Labour (Confederazione Generale Italiana del Lavoro, or CGIL), controlled in effect by the PDS. Employers' groups and the great state bureaucracies also form important pressure groups in this often sharply polarized society.

Italy is part of the European Union and the Council of Europe and belongs to many other international organizations. With its strategic geographic position on the southern flank of Europe, Italy has since World War II played a fairly important role in the North Atlantic Treaty Organization (NATO). (Li.L./R.L.K.)

This article treats the physical and human geography of Italy and its history. For discussion of the major cities of Italy, see the *Macropædia* articles FLORENCE, MILAN, NAPLES, ROME, and VENICE.

The article is divided into the following sections:

PHYSICAL AND HUMAN GEOGRAPHY

The land

RELIEF

Italy is largely mountainous, with 35 percent of its territory occupied by ranges that are higher than 2,300 feet (702 metres). There are two mountain systems: the scenic Alps, parts of which lie within the neighbouring countries of France, Switzerland, Austria, and Croatia; and the Apennines, which form the spine of the entire peninsula and of the island of Sicily. A third mountain system exists in the two large islands to the west, Italian Sardinia and French Corsica.

Mountain ranges. The Alps run in a broad west-to-east arc from the Cadibona Pass, near Savona on the Gulf of Genoa, to the north of Trieste, at the head of the Adriatic Sea. The section properly called Alpine is the border district that includes the highest masses, made up of weathered Hercynian rocks, dating from the Carboniferous or the Permian periods (360 to 245 million years ago). The Alps have rugged, very high peaks, reaching more than 12,800 feet in various spectacular formations, characterized as pyramidal, pinnacled, rounded, or needlelike. The valleys were heavily scoured by glaciers in the Quaternary Period (the past 1.6 million years); there are still more than 1,000 glaciers left, though in a phase of retreat, more than 100 having disappeared in the past half century or so.

The three main Alpine groups

The Alpine mountain mass falls into three main groups: the Western Alps, running north to south in Italy from Aosta to the Cadibona Pass, with Mount Gran Paradiso (13,323 feet) and Mount Viso (12,602 feet); the Central Alps, running west to east from the Western Alps to the Brenner Pass, leading into Austria and the Alto-Adige Valley, also with high peaks, such as Mont Blanc (15,771 feet), the Matterhorn (Italian: Monte Cervino; 14,692 feet), Mount Rosa (Italy's highest peak, shared with Switzerland; 15,203 feet), and Mount Ortles (12,812 feet); and the Eastern Alps, running west to east from the Brenner Pass to Trieste and including the Dolomites (Alpi Dolomitiche), with Mount Marmolada (10,968 feet). The Italian foothills of the Alps, which reach no higher than 8,200 feet, lie between these great ranges and the Po Valley. They are composed mainly of limestone and sedimentary rocks. A notable feature is the karst system of underground caves and streams that are especially characteristic of the Carso, the limestone plateau between the Eastern Alps and Illyria.

The Apennines are the long system of mountains and hills that run down the Italian peninsula from the Cadibona Pass to the tip of Calabria and continue in the island of Sicily. The range is about 1,245 miles long; it is only about 20 miles wide at either end but about 120 miles wide in the Central Apennines, east of Rome, where the Great Rock of Italy (Gran Sasso d'Italia) contains the highest peak (9,554 feet) and the only glacier on the peninsula, Calderone, the southernmost in Europe. The Apennines are predominantly of sandstone and limestone marl (clay) in the north; of limestone and dolomite (magnesium limestone) in the centre; and of limestone, weathered rock, and Hercynian granite in the south. On either side of the central mass are grouped two considerably lower masses, composed in general of more recent and softer rocks, such as sandstone. These are the sub-Apennines, which run in the east from Monferrato to the Gulf of Taranto and in the west from Florence southward through Tuscany and Umbria to Rome. This latter range is separated from the main Apennines by the valleys of the Arno and the Tiber rivers. At the outer flanks of the sub-Apennines two allied series of limestone and volcanic rock extend to the coast. They include, on the west, the Apuan Alps, which are famous for their marbles; farther south, the Metallifere Mountains (more than 3,380 feet), abundant in minerals; then various extinct volcanoes occupied by crater lakes, such as that of Bolsena; then cavernous mountains, such as Lepini and Circeo, and the partially or still fully active volcanic group of the Flegrei Plain and Vesuvius; and finally the limestone mountains of the peninsulas of Amalfi and Cilento. The extensions on the Adriatic coast are simpler, comprising only the small promontory of Mount Conero, the higher peninsula of Gargano (3,465 feet), and the Salentine Peninsula in Puglia. All of these are limestone.

In Sardinia there are two mountain masses, separated by the long plain of Campidano, which runs from the Gulf of Asinara southeastward across the island to the Gulf of Cagliari. The group in the southwest is small and low, formed from sediments, mostly mineralized, perhaps of the early Paleozoic Era (540 million to 245 million years ago). The northeastern mass reaches a height of more than 6,000 feet at Gennargentu; the underlying foundation is basically metamorphic (heat- and pressure-altered) rock, and it is covered in the northeast by Paleozoic granite and partially covered in the northwest by Mesozoic limestones (66.4 million to 245 million years old) and by sandstone and Tertiary Period (1.6 to 66.4 million years old) clays. There are caves on the seacoast and inland where limestones predominate.

Recent volcanic action has produced the Flegrei Plain, near Naples, and the neighbouring islands, such as Ischia; Vesuvius; the Eolie and Lipari islands; and Mount Etna, which is on the island of Sicily. Phenomena that are related to volcanism include thermal springs in the Euganean Hills, *vulcanelli* (mud springs) at Viterbo, and emissions of gas at Pozzuoli.

Seismic activity, leading to earthquakes, is rare in the Alps and the Po Valley; it is infrequent but occasionally strong in the Alpine foothills; and it may be catastrophic in the central and southern Apennines (as in 1980) and on Sicily.

The plains. Plains cover less than a quarter of the area of Italy. Some of these, such as the Po Valley and the Apulian Plain, are ancient sea gulfs filled by alluvium. Others, such as the Plain of Lecce in Puglia, flank the sea on rocky plateaus about 65 to 100 feet high and are formed of ancient land leveled by the sea and subsequently uplifted. Plains in the interior, such as the long Chiana Valley, are made by alluvial or other filling of ancient basins. The most extensive and important plain in Italy, that of the Po Valley, occupies more than 17,000 of the 27,000 square miles of Italian plain land. It ranges in altitude from sea level up to 1,800 feet, the greater part being below 330 feet. Through it run the Po River and all its tributaries and the Reno, Adige, Piave, and Tagliamento rivers. The plain falls into several natural divisions. At its highest end, by the Alpine foothills, it is made up of parallel *ferretto* (red loam composed of ferrous clay) ridges, running from north to south, with areas of gravel and permeable sand between them. This section of the plain is terraced and unproductive, although the rainfall is high. Below this is a section where the rivers rise, their waters eventually providing vital irrigation both for the *marcite* (winter pastures) and for the intensive agriculture of the fertile lower plain. Other notable plains include the *maremme* of Tuscany and Latium, reclaimed marshland

Plain of the Po Valley

with dunes at the edge of the sea; the Agrò Marshes, a recently reclaimed seaward extension of the Roman countryside (*campagna*); the fertile Plain of Campania around Vesuvius; and the rather arid Apulian Plain. In Sicily the Plain of Catania is a good area for growing citrus fruit.

Coastal areas. The seacoasts are quite varied. Along the two Ligurian rivieras, on either side of Genoa, the coast alternates in rapid succession between high, rocky zones and level gravel. From Tuscany to Campania there are long, sandy, crescent beaches and abundant dunes, which are separated by rocky eminences. The coast of Calabria is high and rocky, though sometimes broken by short beaches. The coast of Puglia and, indeed, most of the Adriatic coast is level, although it is dominated by terraced gradients. The majestic delta of the Po, extending from Rimini to Monfalcone, is riddled with the lagoons that are familiar to visitors to Venice. The Carso, the limestone coastal region between Trieste and Istria, is rocky.

DRAINAGE AND SOILS

Rivers. Italian rivers are comparatively short; the longest, the Po, is merely 400 miles long. While three major rivers flow into the Ionian Sea, Puglia has only two rivers flowing to the Adriatic. Along the Adriatic coast a good number run parallel like the teeth of a comb down from the Apennines through Molise, Abruzzi, and Marche regions. The rivers that flow into the Tyrrhenian Sea are longer and more complex and carry greater quantities of water. These include the Volturno, in Campania; the Roman Tiber; and the Arno, which flows through Florence and Pisa. The rivers of the Ligurian rivieras are mainly short and swift-flowing; a few are important simply because cities, such as Genoa, or bathing resorts, such as Rapallo, are built on their deltas. But the prince of Italian rivers is the Po. Rising in the Mount Viso area, it runs across the Plain of Lombardy, through various important cities, such as Turin and Cremona, and is steadily enlarged by the numerous tributaries that join it, especially on its left bank. The Po debouches south of Venice, forming a large delta. In the Veneto there are also rivers that are not tributaries of the Po. One of these, the Adige, at 254 miles the second longest river in Italy, flows through Verona and debouches near Adria, south of Venice. The rivers in the south have imposing floods during winter storms, and those that run through zones of impermeable rock may become dangerous; yet during the summer many of these rivers are completely dry. The rivers of the centre and north are dry in the winter because their headwaters are frozen, but they become full in the spring from melting snow and in the autumn from rainfall.

Lakes. There are about 1,500 lakes in Italy. The most common type is the small, elevated Alpine lake formed by Quaternary glacial excavation during the last 25,000 years. These are of major importance for hydroelectric schemes. Other lakes, such as Bolsena and Albano, in Latium, occupy the craters of extinct volcanoes. There are also coastal lakes, such as those of Lesina and Varano, in Puglia, and lakes resulting from prehistoric faulting, such as Lake Alleghe, near Belluno. The best-known, largest, and most important of the Italian lakes are those cut into valleys of the Alpine foothills by Quaternary glaciers. These, listed in order of size, are Lake Garda, Lake Maggiore, and the lakes of Como, Iseo, and Lugano. They have a semi-Mediterranean climate and are surrounded by groves of olive and citrus trees. Italy also has considerable areas in which, as a result of porous rock, the water systems run underground, forming subterranean streams, sinkholes, and lakes. These are often associated with caves, the most famous of which are those of Castellana, in Puglia.

The soils. Varying climatic conditions in successive eras and differences in altitude and in types of rock have combined to produce in Italy a wide range of soils. Very common is dark-brown podzol, typical in mountains with a lot of flint, where the rainfall is heavy, as in the Alps above about 300 feet. In the Apennines, brown podzolic soils predominate, supporting forests and meadows and pastures. Brown Mediterranean soils also are characteristic of the Apennines and are suitable for agriculture. Rendzinas, typically humus-carbonates, are characteristic of limestone and magnesium limestone mountain pastures and of many meadows and beech forests of the Apennines. Red earth—the famous terra rossa, derived from the residue of limestone rocks—is found mainly in the extreme south, especially in Puglia and southeastern Sicily, where it is the usual soil in vineyards, olive groves, and gardens. Sparse rocky earth, clays, dune sands, and gravel are found in the high mountains, in some volcanic zones, and in gullies in the sub-Apennines. There is also a red loam, or *ferretto*, composed of ferrous (iron) clay.

CLIMATE

Geographically, Italy lies in the temperate zone. Because of the considerable length of the peninsula, there is a variation between the climate of the north, attached to the European continent, and that of the south, surrounded by the Mediterranean. The Alps are a partial barrier against westerly and northerly winds, while both the Apennines and the great plain of northern Italy produce special climatic variations. Sardinia is subject to Atlantic, and Sicily to African, winds. In general, four meteorological situations dominate the Italian climate: the Mediterranean winter cyclone, with a corresponding summer anticyclone; the Alpine summer cyclone, with a consequent winter anticyclone; the Atlantic autumnal cyclone; and the eastern Siberian autumnal anticyclone. The meeting of the two last-mentioned air masses brings heavy and sometimes disastrous rains in the autumn.

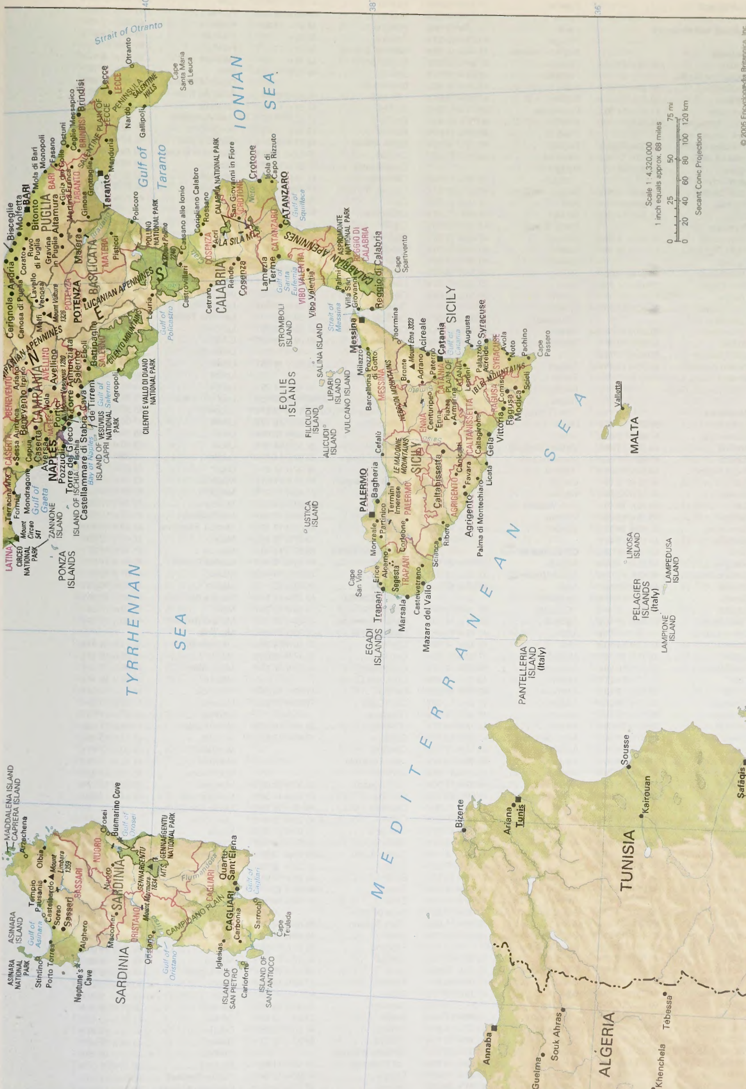
The seven main climatic zones

Italy can be divided into seven main climatic zones. In the most northerly, the Alpine zone, which has a continental mountain climate, temperatures are lower and rainfall higher in the east than in the west. The average temperature at Bardonecchia, in the west, is 45.3° F (7.4° C), and the rainfall is 26 inches (660 millimetres); at Cortina d'Ampezzo, in the east, the figures are 43.9° F (6.6° C) and 41.5 inches. In the Valle d'Aosta in the west, the permanent snow line is at 10,200 feet, but in the Julian Alps it is as low as 8,350 feet. In autumn and in late winter the hot, dry wind that is known as the foehn blows from Switzerland or Austria, and in the east the cold, dry bora blows with gusts of up to 125 miles per hour (200 kilometres per hour). Rain falls in the summer in the higher and more remote areas and in the spring and autumn at the periphery. Snow falls only in the winter, but the snowfall varies from about 10 to 33 feet in different years and in relation to altitude or proximity to the sea. More snow falls in the foothills than in the mountains and more in

© M. Fritz/Superstock

Lake Garda, with the town of Malcesine on its eastern shore, northern Italy.


Cities over 1,000,000
Cities 200,000 to 1,000,000
Cities 50,000 to 200,000
Cities under 50,000
Other localities
National capitals
Provincial capitals
BARI Regional capitals
MOLISE Regional names
ROME Provincial names
International boundaries
Regional boundaries
Provincial boundaries
Canals
Dams
Glaciers
National parks
Historical sites
Other points of interest
Spot elevations in metres (1 m = 3.28 ft)
GERMANY
AUSTRIA
SWITZERLAND
LIECHTENSTEIN
HUNGARY
SLOVENIA
CROATIA
BOSNIA AND HERZEGOVINA
SERBIA AND MONTENEGRO
FRANCE
MONACO
CORSICA (Fr.)
ADRIATIC SEA
LIGURIAN SEA
TUSCANY ARCHIPELAGO
Gulf of Venice
Gulf of Genoa
ALPS
APENNINES
PIEDMONT
LOMBARDY
VENETO
LIGURIA
EMILIA-ROMAGNA
TUSCANY
UMBRIA
MARCHE
LATIUM
ABRUZZI
MOLISE
ROME
VATICAN CITY
MILAN
TURIN
GENOA
VENICE
FLORENCE
BOLOGNA
TRIESTE
ANCONA
PERUGIA
L'AQUILA
CAMPOBASSO
SAN MARINO


TYRRHENIAN SEA
IONIAN SEA
MEDITERRANEAN SEA
Strait of Otranto
Gulf of Taranto
SARDINIA
SICILY
CALABRIA
BASILICATA
PUGLIA
CAMPANIA
TUNISIA
ALGERIA
MALTA
Valletta
Tunis
Bizerte
Ariana
Sousse
Kairouan
Şafāqis
Annaba
Souk Ahras
Guelma
Tebessa
Khenchela
PALERMO
Messina
Catania
Syracuse
Reggio di Calabria
CATANZARO
Crotone
Cosenza
POTENZA
Matera
BARI
Taranto
Brindisi
Lecce
Otranto
NAPLES
Salerno
Caserta
Avellino
Benevento
CAGLIARI
Sassari
Nuoro
Oristano
Olbia
Alghero
Iglesias
Trapani
Marsala
Agrigento
Caltanissetta
Enna
Ragusa
Gela
EOLIE ISLANDS
EGADI ISLANDS
USTICA ISLAND
PANTELLERIA ISLAND (Italy)
PELAGIER ISLANDS (Italy)
LAMPEDUSA ISLAND
LINOSA ISLAND
LAMPIONE ISLAND
PONZA ISLANDS
ISLAND OF CAPRI
ISLAND OF ISCHIA
ISLAND OF SAN PIETRO
ISLAND OF SANT'ANTIOCO
ASINARA ISLAND
MADDALENA ISLAND
CAPRERA ISLAND
Mount Etna 3323
Mount Vesuvius 1280
Scale 1 : 4,320,000
1 inch equals approx. 68 miles
0 25 50 75 mi
0 20 40 60 80 100 120 km
Secant Conic Projection
© 2005 Encyclopædia Britannica, Inc.

MAP INDEX

Political subdivisions

Regions
Abruzzi 42 15 N 13 45 E
Apulia, see Puglia
Basilicata 40 30 N 16 30 E
Calabria 39 00 N 16 30 E
Campania 41 00 N 14 30 E
Emilia-Romagna 44 45 N 11 00 E
Friuli-Venezia Giulia (Venezia Giulia) 46 00 N 13 00 E
Latium (Lazio) 42 00 N 12 30 E
Liguria 44 30 N 8 50 E
Lombardy (Lombardia) 45 40 N 9 30 E
Marche 43 30 N 13 15 E
Molise 41 38 N 14 29 E
Piedmont (Piemonte) 45 00 N 8 00 E
Puglia (Apulia) 41 15 N 16 15 E
Sardinia (Sardegna) 40 00 N 9 00 E
Sicily (Sicilia) 37 45 N 14 15 E
Trentino-Alto Adige 46 30 N 11 20 E
Tuscany (Toscana) 43 25 N 11 00 E
Umbria 43 00 N 12 30 E
Valle d'Aosta 45 45 N 7 15 E
Veneto (Venetia) 45 30 N 11 45 E
Venezia Giulia, see Friuli-Venezia Giulia

Provinces
Agrigento 37 27 N 13 30 E
Alessandria 44 49 N 8 42 E
Ancona 43 38 N 13 30 E
Arezzo 43 32 N 11 50 E
Ascoli Piceno 43 01 N 13 34 E
Asti 44 55 N 8 10 E
Avellino 40 59 N 15 09 E
Bari 40 56 N 16 40 E
Belluno 46 15 N 12 08 E
Benevento 41 15 N 14 17 E
Bergamo 45 50 N 9 48 E
Biella 45 35 N 8 05 E
Bologna 44 28 N 11 26 E
Bolzano-Bozen 46 43 N 11 30 E
Brescia 45 38 N 10 18 E
Brindisi 40 35 N 17 40 E
Cagliari 39 30 N 8 45 E
Caltanissetta 37 24 N 14 10 E
Campobasso 41 38 N 14 35 E
Caserta 41 14 N 14 10 E
Catania 37 23 N 14 40 E
Catanzaro 38 55 N 16 26 E
Chieti 42 07 N 14 21 E
Como 45 59 N 9 13 E
Cosenza 39 28 N 16 25 E
Cremona 45 14 N 9 56 E
Crotone 39 10 N 17 00 E
Cuneo 44 31 N 7 34 E
Enna 37 35 N 14 26 E
Ferrara 44 48 N 11 50 E
Florence (Firenze) 43 50 N 11 20 E
Foggia 41 27 N 15 32 E
Forlì 44 05 N 12 02 E
Frosinone 41 37 N 13 27 E
Genoa (Genova) 44 30 N 9 04 E
Gorizia 45 55 N 13 30 E
Grosseto 42 50 N 11 15 E
Imperia 43 58 N 7 47 E
Isernia 41 36 N 14 14 E
La Spezia 44 15 N 9 42 E
L'Aquila 42 05 N 13 40 E
Latina 41 27 N 13 06 E
Lecce 40 13 N 18 10 E
Lecco 45 55 N 9 25 E
Leghorn (Livorno) 43 14 N 10 35 E
Lodi 45 15 N 9 30 E
Lucca 44 02 N 10 27 E
Macerata 43 12 N 13 10 E
Mantua (Mantova) 45 10 N 10 47 E
Massa-Carrara 44 15 N 10 03 E
Matera 40 30 N 16 25 E
Messina 38 03 N 14 52 E
Milan (Milano) 45 30 N 9 30 E
Modena 44 30 N 10 54 E
Naples (Napoli) 40 53 N 14 25 E
Novara 45 58 N 8 24 E
Nuoro 40 10 N 9 20 E
Oristano 39 54 N 8 36 E
Padua (Padova) 45 21 N 11 49 E
Palermo 37 49 N 13 35 E
Parma 44 42 N 10 05 E
Pavia 45 07 N 9 08 E
Perugia 43 03 N 12 33 E
Pesaro e Urbino 43 40 N 12 38 E
Pescara 42 20 N 13 57 E
Piacenza 44 53 N 9 35 E
Pisa 43 25 N 10 43 E
Pistoia 43 58 N 10 50 E
Pordenone 45 57 N 12 39 E
Potenza 40 35 N 15 44 E
Prato 43 58 N 11 05 E
Ragusa 36 55 N 14 36 E
Ravenna 44 25 N 11 59 E
Reggio di Calabria 38 19 N 16 05 E
Reggio nell'Emilia 44 37 N 10 37 E
Rieti 42 18 N 12 52 E
Rimini 44 00 N 12 35 E
Rome (Roma) 41 58 N 12 40 E
Rovigo 45 02 N 11 50 E
Salerno 40 27 N 15 16 E
Sassari 40 40 N 9 00 E
Savona 44 18 N 8 16 E
Siena 43 13 N 11 24 E
Sondrio 46 10 N 10 03 E
Syracuse (Siracusa) 37 03 N 15 00 E
Taranto 40 37 N 17 15 E
Teramo 42 39 N 13 41 E
Terni 42 41 N 12 19 E
Trapani 37 50 N 12 40 E
Trento 46 08 N 11 07 E
Treviso 45 50 N 12 13 E
Trieste 45 20 N 13 25 E
Turin (Torino) 45 08 N 7 22 E
Udine 46 10 N 13 00 E
Varese 45 48 N 8 48 E
Venice (Venezia) 45 35 N 12 34 E
Verbano-Cusio-Ossola 46 05 N 8 20 E
Vercelli 45 37 N 8 10 E
Verona 45 25 N 11 02 E
Vibo Valentia 38 40 N 16 10 E
Vicenza 45 40 N 11 27 E
Viterbo 42 25 N 12 05 E

Cities and towns
Acireale 37 37 N 15 10 E
Adrano (Aderno) 37 40 N 14 50 E
Adria 45 03 N 12 03 E
Agrigento (Girgenti) 37 19 N 13 34 E
Agropoli 40 21 N 14 59 E
Alatri 41 43 N 13 21 E
Alba 44 42 N 8 02 E
Alcamo 37 59 N 12 58 E
Alessandria 44 54 N 8 37 E
Alghero 40 33 N 8 19 E
Altamura 40 49 N 16 33 E
Anagni 41 44 N 13 09 E
Ancona 43 38 N 13 30 E
Andria 41 13 N 16 17 E
Anzio (Antium) 41 27 N 12 37 E
Aosta 45 44 N 7 20 E
Aquileia 45 46 N 13 22 E
Arezzo 43 25 N 11 53 E
Argenta 44 37 N 11 50 E
Ariano Irpino 41 09 N 15 05 E
Arpino 41 39 N 13 36 E
Arzachena 41 05 N 9 23 E
Ascoli Piceno 42 51 N 13 34 E
Assisi 43 04 N 12 37 E
Asti 44 54 N 8 12 E
Atri 42 35 N 13 58 E
Augusta 37 13 N 15 13 E
Avellino 40 54 N 14 47 E
Aversa 40 58 N 14 12 E
Avezzano 42 02 N 13 25 E
Avola 36 54 N 15 08 E
Bagheria 38 05 N 13 30 E
Barcellona Pozzo di Gotto 38 09 N 15 13 E
Bardonecchia 45 05 N 6 42 E
Bari 41 08 N 16 51 E
Barletta 41 19 N 16 17 E
Bassano del Grappa 45 46 N 11 44 E
Battipaglia 40 37 N 14 58 E
Belgioioso 45 10 N 9 19 E
Belluno 46 09 N 12 13 E
Benevento 41 08 N 14 45 E
Bergamo 45 41 N 9 43 E
Biella 45 34 N 8 03 E
Bisceglie 41 14 N 16 30 E
Bitonto 41 06 N 16 41 E
Bobbio 44 46 N 9 23 E
Bologna 44 29 N 11 20 E
Bolsena 42 39 N 11 59 E
Bolzano (Bozen) 46 31 N 11 22 E
Bordighera 43 46 N 7 39 E
Brescia 45 33 N 10 15 E
Bressanone (Brixen) 46 43 N 11 39 E
Brindisi 40 38 N 17 56 E
Bronte 37 47 N 14 50 E
Burano 45 29 N 12 25 E
Busto Arsizio 45 37 N 8 51 E
Cagliari 39 13 N 9 07 E
Caltanissetta 37 29 N 14 04 E
Campobasso 41 34 N 14 39 E
Canosa di Puglia 41 13 N 16 04 E
Cantù 45 44 N 9 08 E
Capannori 43 50 N 10 34 E
Capua 41 06 N 14 12 E
Carbonia 39 10 N 8 31 E
Carloforte 39 08 N 8 18 E
Carpi 44 47 N 10 53 E
Carrara 44 05 N 10 06 E
Casale Monferrato 45 08 N 8 27 E
Cascina 43 41 N 10 33 E
Caserta 41 04 N 14 20 E
Cassino 41 30 N 13 49 E
Castel Gandolfo 41 45 N 12 39 E
Castelfidardo 43 28 N 13 33 E
Castelfranco Veneto 45 40 N 11 55 E
Castellammare di Stabia 40 42 N 14 29 E
Castelsardo 40 55 N 8 43 E
Castelvetrano 37 41 N 12 47 E
Castrovillari 39 49 N 16 12 E
Catania 37 30 N 15 06 E
Catanzaro 38 54 N 16 35 E
Cava de' Tirreni 40 42 N 14 42 E
Cefalù 38 02 N 14 01 E
Ceglie Messapico 40 39 N 17 31 E
Cento 44 43 N 11 17 E
Centuripe 37 37 N 14 44 E
Cerignola 41 16 N 15 54 E
Cesena 44 08 N 12 15 E
Cetraro 39 31 N 15 56 E
Chianciano Terme 43 02 N 11 49 E
Chiavari 44 19 N 9 19 E
Chieti 42 21 N 14 10 E
Chioggia 45 13 N 12 17 E
Città di Castello 43 27 N 12 14 E
Cividale del Friuli 46 06 N 13 25 E
Civita Castellana 42 17 N 12 25 E
Civitavecchia 42 06 N 11 48 E
Comiso 36 56 N 14 36 E
Como 45 47 N 9 05 E
Conegliano 45 53 N 12 18 E
Coni, see Cuneo
Corato 41 09 N 16 25 E
Cori (Cora) 41 39 N 12 55 E
Corigliano Calabro 39 36 N 16 31 E
Corleone 37 49 N 13 18 E
Corneto, see Tarquinia
Cortina d'Ampezzo 46 32 N 12 08 E
Cortona 43 16 N 11 59 E
Cosenza 39 18 N 16 15 E
Crema 45 22 N 9 41 E
Cremona 45 07 N 10 02 E
Crotone 39 05 N 17 08 E
Cuneo (Coni) 44 23 N 7 32 E
Desio 45 37 N 9 13 E
Eboli 40 36 N 15 04 E
Empoli 43 43 N 10 57 E
Enna 37 34 N 14 16 E
Erice (Monte San Giuliano) 38 02 N 12 35 E
Este 45 14 N 11 39 E
Fabriano 43 20 N 12 54 E
Faenza 44 17 N 11 53 E
Fano 43 50 N 13 01 E
Favara 37 19 N 13 39 E
Feltre 46 01 N 11 54 E
Ferentino 41 42 N 13 15 E
Fermo 43 09 N 13 43 E
Ferrara 44 50 N 11 35 E
Fidenza 44 52 N 10 03 E
Florence (Firenze or Florentia) 43 46 N 11 15 E
Foggia 41 27 N 15 34 E
Foligno 42 57 N 12 42 E
Fondi 41 21 N 13 25 E
Forlì 44 13 N 12 03 E
Formia 41 15 N 13 37 E
Fossano 44 33 N 7 43 E
Frascati 41 48 N 12 41 E
Frosinone 41 38 N 13 19 E
Gallarate 45 40 N 8 47 E
Gela 37 04 N 14 15 E
Genoa (Genova) 44 25 N 8 57 E
Gioia del Colle 40 48 N 16 55 E
Girgenti, see Agrigento
Gorgonzola 45 32 N 9 24 E
Gorizia (Gorz) 45 57 N 13 38 E
Gravina in Puglia 40 49 N 16 25 E
Grosseto 42 46 N 11 08 E
Grottaglie 40 32 N 17 26 E
Guastalla 44 55 N 10 39 E
Gubbio 43 21 N 12 35 E
Iesi, see Jesi
Iglesias 39 19 N 8 32 E
Imola 44 21 N 11 42 E
Imperia 43 53 N 8 03 E
Ischia 40 44 N 13 57 E
Isernia 41 36 N 14 14 E
Ivrea 45 28 N 7 52 E
Jesi (Iesi) 43 31 N 13 14 E
La Spezia 44 07 N 9 50 E
Lamezia Terme 38 55 N 16 15 E
Lanciano 42 14 N 14 23 E
L'Aquila 42 22 N 13 22 E
Larderello 43 14 N 10 53 E
Latina 41 28 N 12 52 E
Lecce 40 23 N 18 11 E
Lecco 45 51 N 9 23 E
Leghorn (Livorno) 43 33 N 10 19 E
Legnano 45 36 N 8 54 E
Licata 37 06 N 13 56 E
Lodi 45 19 N 9 30 E
Lucca 43 50 N 10 29 E
Lucera 41 30 N 15 20 E
Lugo 44 25 N 11 54 E
Macerata 43 18 N 13 27 E
Magenta 45 28 N 8 53 E
Manduria 40 24 N 17 38 E
Manfredonia 41 38 N 15 55 E
Mantua (Mantova) 45 09 N 10 48 E
Marino 41 46 N 12 39 E
Marsala 37 48 N 12 26 E
Martina Franca 40 42 N 17 20 E
Massa 44 01 N 10 09 E
Matera 40 40 N 16 36 E
Mazara del Vallo 37 39 N 12 35 E
Melfi 41 00 N 15 39 E
Merano (Meran) 46 40 N 11 09 E
Messina (Messana) 38 11 N 15 34 E
Mestre 45 29 N 12 15 E
Milan (Milano) 45 28 N 9 12 E
Milazzo 38 13 N 15 14 E
Mirandola 44 53 N 11 04 E
Modena 44 40 N 10 55 E
Modica 36 52 N 14 46 E
Mola di Bari 41 04 N 17 05 E
Molfetta 41 12 N 16 36 E
Moncalieri 45 00 N 7 41 E
Mondovì 44 23 N 7 49 E
Monfalcone 45 49 N 13 32 E
Monreale 38 05 N 13 17 E
Montagnana 45 14 N 11 28 E
Monte San Giuliano, see Erice
Monte Sant'Angelo 41 42 N 15 57 E
Montecatini Terme 43 53 N 10 46 E
Monza 45 35 N 9 16 E
Naples (Napoli or Neapolis) 40 50 N 14 15 E
Nardò 40 11 N 18 02 E
Narni 42 31 N 12 31 E
Nola 40 55 N 14 33 E
Noto 36 53 N 15 04 E
Novara 45 28 N 8 38 E
Novi Ligure 44 46 N 8 47 E
Nuoro 40 19 N 9 20 E
Olbia 40 55 N 9 31 E
Oristano 39 54 N 8 36 E
Orosei 40 23 N 9 42 E
Ortona 42 21 N 14 24 E
Orvieto 42 43 N 12 07 E
Otranto 40 09 N 18 30 E
Pachino 36 43 N 15 05 E
Padua (Padova) 45 25 N 11 53 E
Palazzolo Acreide 37 04 N 14 54 E
Palermo 38 07 N 13 22 E

Parma 44 48 N 10 20 E
Paternò 37 34 N 14 54 E
Pavia 45 10 N 9 10 E
Perugia (Perusia) 43 08 N 12 22 E
Pesaro (Pisaurum) 43 54 N 12 55 E
Pescara 42 28 N 14 13 E
Peschiera del Garda 45 26 N 10 42 E
Piacenza 45 01 N 9 40 E
Piazza Armerina 37 23 N 14 22 E
Pinerolo 44 53 N 7 21 E
Piombino 42 55 N 10 32 E
Pisa 43 43 N 10 23 E
Pisaurum, see Pesaro
Pistoia 43 55 N 10 54 E
Pordenone 45 57 N 12 39 E
Portici 40 49 N 14 20 E
Porto Torres 40 50 N 8 24 E
Portofino 44 18 N 9 12 E
Portogruaro 45 47 N 12 50 E
Potenza 40 38 N 15 48 E
Pozzuoli 40 49 N 14 07 E
Prato 43 53 N 11 06 E
Ragusa 36 55 N 14 44 E
Rapallo 44 21 N 9 14 E
Ravenna 44 25 N 12 12 E
Recanati 43 24 N 13 32 E
Reggio di Calabria 38 06 N 15 39 E
Reggio nell'Emilia 44 43 N 10 36 E
Rende 39 20 N 16 11 E
Rieti 42 24 N 12 51 E
Rimini 44 04 N 12 34 E
Rivoli 45 04 N 7 31 E
Rome (Roma) 41 54 N 12 29 E
Rossano 39 34 N 16 38 E
Rovereto 45 53 N 11 02 E
Rovigo 45 04 N 11 47 E
Ruvo di Puglia 41 07 N 16 29 E
Salerno 40 41 N 14 47 E
Saluzzo 44 39 N 7 29 E
San Gimignano 43 28 N 11 02 E
San Giovanni Rotondo 41 42 N 15 44 E
San Giovanni Valdarno 43 34 N 11 32 E
San Giuliano Terme 43 46 N 10 26 E
San Remo 43 49 N 7 46 E
San Severo 41 41 N 15 23 E
Sansepolcro 43 34 N 12 08 E
Sarroch 39 04 N 9 00 E
Sarzana 44 07 N 9 58 E
Sassari 40 43 N 8 34 E
Savona 44 17 N 8 30 E
Schio 45 43 N 11 21 E
Sciacca 37 31 N 13 03 E
Scicli 36 47 N 14 42 E
Senigallia (Sena Gallica or Sinigaglia) 43 43 N 13 13 E
Sessa Aurunca 41 14 N 13 56 E
Sesto San Giovanni 45 32 N 9 14 E
Siena 43 19 N 11 21 E
Siracusa, see Syracuse
Sondrio 46 10 N 9 52 E
Sora 41 43 N 13 37 E
Spezia, see La Spezia
Spoleto 42 44 N 12 44 E
Stintino 40 56 N 8 13 E
Stresa 45 53 N 8 32 E
Subiaco 41 55 N 13 06 E
Sulmona 42 03 N 13 55 E
Syracuse (Siracusa) 37 04 N 15 18 E
Taormina 37 51 N 15 17 E
Taranto (Taras or Tarentum) 40 28 N 17 14 E
Tarquinia (Corneto) 42 15 N 11 45 E
Tempio Pausania 40 54 N 9 06 E
Teramo 42 39 N 13 42 E
Termini Imerese 37 59 N 13 42 E
Terni 42 34 N 12 37 E
Terracina 41 17 N 13 15 E
Tivoli (Tibur) 41 58 N 12 48 E
Todi 42 47 N 12 24 E
Torino, see Turin
Torre Annunziata 40 45 N 14 27 E
Torre del Greco 40 47 N 14 22 E
Tortona 44 54 N 8 52 E
Toscanella, see Tuscania
Trapani 38 01 N 12 29 E
Trento 46 04 N 11 08 E
Treviso 45 40 N 12 15 E
Trieste 45 40 N 13 46 E
Turin (Torino) 45 03 N 7 40 E
Tuscania (Toscanella) 42 25 N 11 52 E
Udine 46 03 N 13 14 E
Umbertide 43 18 N 12 20 E
Urbino 43 43 N 12 38 E
Vallombrosa 43 44 N 11 32 E
Varallo 45 49 N 8 15 E
Varese 45 48 N 8 50 E
Vasto 42 07 N 14 42 E
Venice (Venezia) 45 27 N 12 21 E
Venosa (Venusia) 40 58 N 15 49 E
Ventimiglia 43 47 N 7 36 E
Verbania 45 56 N 8 33 E
Vercelli 45 19 N 8 25 E
Verona 45 27 N 11 00 E
Viareggio 43 52 N 10 14 E
Vibo Valentia 38 40 N 16 06 E
Vicenza 45 33 N 11 33 E
Vigevano 45 19 N 8 51 E
Villa San Giovanni 38 13 N 15 38 E
Viterbo 42 25 N 12 06 E
Vittoria 36 57 N 14 32 E
Vittorio Veneto 45 59 N 12 18 E
Voghera 44 59 N 9 01 E
Volterra 43 24 N 10 51 E

Physical features and points of interest

Abruzzi (Abruzzese) Apennines, *mountains* 42 00 N 14 00 E
Abruzzi National Park 41 45 N 13 50 E
Adamello Mountains 46 10 N 10 35 E
Adamello Regional Nature Park 46 20 N 10 10 E
Adda, *river* 45 08 N 9 53 E
Adige (Athesis or Etsch), *river* 45 10 N 12 20 E
Adriatic Sea 43 00 N 16 00 E
Aegadian, see Egadi Islands
Albano, Lake 41 45 N 12 40 E
Alicudi Island 38 32 N 14 21 E
Alps, *mountains* 46 25 N 10 00 E
Amiata, Mount 42 53 N 11 37 E
Apennines (Appennino), *mountains* 43 00 N 13 00 E
Apuan (Apuane) Alps, *mountains* 44 04 N 10 17 E
Apuan (Apuane) Alps Regional Nature Park 44 03 N 10 20 E
Apulian (Puglia) Plain 41 35 N 15 25 E
Argentario, Mount 42 24 N 11 09 E
Arno (Arnus), *river* 43 41 N 10 17 E
Asinara, Gulf of 41 00 N 8 32 E
Asinara Island 41 04 N 8 16 E
Aspromonte National Park 38 20 N 16 05 E
Athesis, see Adige
Aurunci Mountains 41 22 N 13 40 E
Bellunesi Dolonites National Park 46 15 N 12 00 E
Benaco, see Garda, Lake
Bolsena, Lake 42 36 N 11 56 E
Bonifacio, Strait of 41 18 N 9 15 E
Bracciano, Lake 42 07 N 12 14 E
Bradano, *river* 40 23 N 16 51 E
Brenner (Brennero) Pass 47 00 N 11 30 E
Brenta, *river* 45 11 N 12 18 E
Buemarino Cove 40 15 N 9 37 E
Cadibona Pass 44 20 N 8 22 E
Cagliari, Gulf of 39 08 N 9 11 E
Calabrian (Calabro) Apennines, *mountains* 39 00 N 16 30 E
Campanian Apennines, *mountains* 40 50 N 14 45 E
Campidano Plain, *area* 39 30 N 8 47 E
Capraia Island 43 02 N 9 49 E
Caprera Island 41 12 N 9 28 E
Capri, Island of 40 33 N 14 13 E
Carso Plateau 45 40 N 14 00 E
Casentinesi Forest National Park 43 50 N 11 45 E
Catania, Gulf of 37 24 N 15 09 E
Catania, Plain of 37 25 N 14 51 E
Chianti Mountains 43 32 N 11 25 E
Chienti, *river* 43 18 N 13 45 E
Cilento e Vallo di Diano National Park 40 15 N 15 25 E
Cilento Mountains 40 20 N 15 15 E
Cimini Mountains 42 20 N 12 10 E
Cimone, Mount 44 12 N 10 42 E
Circeo, Mount 41 14 N 13 03 E
Clitunno, *river* 42 56 N 12 37 E
Comacchio, Lake 44 38 N 12 06 E
Como (Lario), Lake 46 00 N 9 17 E
Conero, Mount 43 33 N 13 36 E
Cottian (Cozie) Alps, *mountains* 44 45 N 7 00 E
Crati, *river* 39 43 N 16 31 E
Cusio, see Orta, Lake
Delta Padano National Park 44 45 N 12 15 E
Dolomite (Dolomiti or Dolomitiche) Alps, *mountains* 46 23 N 11 51 E
Dora Baltea, *river* 45 11 N 8 03 E
Dora Riparia, *river* 45 05 N 7 44 E
Egadi (Aegadian) Islands 37 58 N 12 16 E
Elba Island 42 46 N 10 17 E
Eolie and Lipari Islands 38 30 N 14 57 E
Ernici Mountains 41 47 N 13 25 E
Esino, *river* 43 39 N 13 22 E
Etna (Mongibello), Mount 37 45 N 15 00 E
Etsch, see Adige
Filicudi Island 38 34 N 14 34 E
Flumendosa, *river* 39 26 N 9 37 E
Foglia, *river* 43 55 N 12 54 E
Fortore, *river* 41 55 N 15 17 E
Gaeta, Gulf of 41 06 N 13 30 E
Garda (Benaco), Lake 45 40 N 10 41 E
Gargano National Park 41 45 N 15 45 E
Gargano Promontory 41 50 N 16 00 E
Gennargentu Mountains 40 01 N 9 19 E
Genoa (Genova), Gulf of 44 10 N 8 55 E
Giannutri Island 42 15 N 11 06 E
Giglio Island 42 21 N 10 54 E
Giulie, see Julian Alps
Gorgona Island 43 26 N 9 54 E
Graian (Graie) Alps 45 30 N 7 10 E
Gran Paradiso, Mount 45 32 N 7 16 E
Gran Paradiso National Park 45 32 N 7 18 E
Great Rock and Laga Mountains National Park 42 30 N 13 30 E
Great Rock of Italy (Gran Sasso d'Italia), *mountains* 42 27 N 13 42 E
Great Saint Bernard (Gran San Bernardo) Pass 45 50 N 7 10 E
Gulf of Orosei, Gennargentu Mountains, and Asinara Island National Park 40 00 N 9 30 E
Iblei Mountains 37 10 N 14 55 E
Ionian Sea 39 00 N 18 30 E
Ischia, Island of 40 43 N 13 54 E
Iseo, Lake 45 43 N 10 04 E
Julian (Giulie) Alps, *mountains* 46 20 N 13 45 E
La Sila Mountains 39 15 N 16 30 E
Lampedusa Island 35 31 N 12 35 E
Lampione Island 35 33 N 12 19 E
Lao, *river* 39 47 N 15 48 E
Lario, see Como, Lake
Le Madonie Mountains 37 52 N 13 58 E
Lecce, Plain of 40 30 N 17 35 E
Lepini Mountains 41 35 N 13 00 E
Lepontine Alps 46 25 N 8 40 E
Lesina Lagoon 41 53 N 15 26 E
Ligurian (Ligure) Apennines, *mountains* 44 30 N 9 00 E
Ligurian (Ligure) Sea 43 30 N 9 00 E
Limbara, Mount 40 51 N 9 10 E
Linosa Island 35 52 N 12 52 E
Lipari, see Eolie and Lipari Islands
Lipari Island 38 29 N 14 56 E
Liri, *river* 41 25 N 13 52 E
Little Saint Bernard (Piccolo San Bernardo) Pass 45 41 N 6 53 E
Livenza, *river* 45 35 N 12 51 E
Lucanian (Lucano) Apennines, *mountains* 40 30 N 16 00 E
Lugano, Lake 45 58 N 9 00 E
Maddalena Island 41 14 N 9 25 E
Maggiorasca, Mount 44 33 N 9 29 E
Maggiore, Lake 45 57 N 8 39 E
Maiella National Park 42 10 N 14 00 E
Manfredonia, Gulf of 41 35 N 16 05 E
Maremma, *region* 42 30 N 11 30 E
Maritime Alps, *mountains* 44 15 N 7 10 E
Marmolada, Mount 46 26 N 11 51 E
Marmora, Mount 39 59 N 9 20 E
Matese Mountains 41 27 N 14 22 E
Mediterranean Sea 35 00 N 12 00 E
Messina, Strait of 38 15 N 15 35 E
Metallifere Mountains 43 11 N 10 56 E
Metauro, *river* 43 50 N 13 03 E
Mincio, *river* 45 04 N 10 59 E
Monferrato Plateau 44 50 N 8 20 E
Mongibello, see Etna, Mount
Montecristo Island 42 20 N 10 19 E
Montone, *river* 44 24 N 12 14 E
Naples (Napoli), Bay of 40 43 N 14 10 E
Nebrodi Mountains 37 54 N 14 35 E
Neptune's Cave (Nettuno, Grotta di) 40 34 N 8 09 E
Ombrone, *river* 42 39 N 11 00 E
Orobie Mountains 46 00 N 10 00 E
Orosei, Gulf of 40 15 N 9 44 E
Orta (Cusio), Lake 45 49 N 8 24 E
Ortles, Mount 46 31 N 10 33 E
Otranto, Strait of 40 00 N 19 00 E
Padus, see Po
Panaro, *river* 44 55 N 11 25 E
Pantelleria Island 36 47 N 12 00 E
Pelagier (Pelagie) Islands 35 40 N 12 40 E
Pennine Alps 46 05 N 7 50 E
Perugia, see Trasimeno, Lake
Pianosa Island 42 35 N 10 04 E
Piave, *river* 45 32 N 12 44 E
Piccolo San Bernardo, see Little Saint Bernard Pass
Po (Padus), *river* 44 57 N 12 04 E
Policastro, Gulf of 40 00 N 15 35 E
Pollino, Mount 39 55 N 16 11 E

Pollino National Park	39 57 N	16 10 E
Pontine (Pontino) Marshes	41 25 N	12 55 E
Ponza (Pontine) Islands	40 55 N	12 57 E
Potenza, *river*	43 25 N	13 40 E
Puglia, see Apulian Plain		
Reno, *river*	44 38 N	12 16 E
Rhaetian (Retiche) Alps	46 30 N	10 00 E
Roia (Roya), *river*	43 48 N	7 35 E
Ronco, *river*	44 24 N	12 12 E
Rosa, Mount	45 55 N	7 53 E
Sabini Mountains	42 13 N	12 50 E
Salerno, Gulf of	40 32 N	14 42 E
Salentine Hills	40 02 N	18 13 E
Salentine (Salentina) Peninsula	40 30 N	18 00 E
Salina Island	38 34 N	14 50 E
Salso, *river*	37 06 N	13 57 E
San Pietro, Island of	39 08 N	8 17 E
Santa Maria di Leuca, Cape	39 47 N	18 22 E
Sant'Antioco, Island of	39 02 N	8 25 E
Sardinia (Sardegna), *island*	40 00 N	9 00 E
Savio, *river*	44 19 N	12 20 E
Secchia, *river*	45 04 N	11 00 E
Segesta, *historical site*	37 56 N	12 50 E
Sesia, *river*	45 05 N	8 37 E
Sibillini Mountains National Park	43 00 N	13 15 E
Sicily (Sicilia), *island*	37 30 N	14 00 E
Sieve, *river*	43 46 N	11 26 E
Simbruini Mountains	41 52 N	13 15 E
Simeto, *river*	37 24 N	15 06 E
Stelvio National Park	46 30 N	10 35 E
Stelvio Pass	46 32 N	10 27 E
Stromboli Island	38 47 N	15 13 E
Tagliamento, *river*	45 38 N	13 06 E
Tanaro, *river*	45 01 N	8 47 E
Taranto, Gulf of	40 08 N	17 20 E
Taro, *river*	45 00 N	10 15 E
Terminillo, Mount	42 28 N	13 00 E
Teulada, Cape	38 52 N	8 38 E
Teverone, see Aniene		
Tiber (Tevere), *river*	41 44 N	12 14 E
Ticino, *river*	45 09 N	9 14 E
Ticino Valley Regional Nature Park	45 20 N	8 45 E
Tirso, *river*	39 53 N	8 32 E
Trasimeno (Perugia), Lake	43 08 N	12 06 E
Trebbia, *river*	45 04 N	9 41 E
Tronto, *river*	42 54 N	13 55 E
Tuscan-Emilian (Tosco-Emiliano) Apennines, *mountains*	44 00 N	11 30 E
Tuscany Archipelago National Park	43 00 N	10 00 E
Tyrrhenian (Tirreno) Sea	40 00 N	12 00 E
Umbrian-Marchigian (Umbro-Marchigiano) Apennines, *mountains*	43 00 N	13 00 E
Ustica Island	38 42 N	13 11 E
Vaiont Dam	46 16 N	12 20 E
Valtellina Valley	46 11 N	9 55 E
Varano, Lake, *lagoon*	41 53 N	15 45 E
Venice (Venezia), Gulf of	45 15 N	13 00 E
Vesuvius, Mount, *volcano*	40 49 N	14 26 E
Viso, Mount	44 40 N	7 07 E
Volturno, *river*	41 01 N	13 55 E
Vulcano Island	38 24 N	14 58 E
Vulture, Mount	40 57 N	15 38 E

the eastern than in the western Alps. Around the lakes the climate is milder, the average temperature in January at Milan being 34° F (1° C), while at Salò, on Lake Garda, it is 39° F (4° C).

The Po Valley has hot summers but severe winters, worse in the interior than toward the eastern coast. At Turin the winter average is 32.5° F (0.3° C) and the summer average 74° F (23° C). Rain falls mainly in the spring and autumn and increases with altitude. There is scant snow, and that falls only on the high plain. The temperatures along the Adriatic coast rise steadily from north to south, partly because of the descending latitude and partly because the prevailing winds are easterly in the north but southerly in the south. The average annual mean temperature rises from 56.5° F (13.6° C) at Venice to 61° F (16° C) at Ancona and 63° F (17° C) at Bari. There is scant rain: Venice has an average of 29.5 inches, Ancona 25.5 inches, and Bari 23.6 inches.

In the Apennines the winters vary in severity according to the altitude. Except at specific locations, there are but moderate amounts of both rain and snow; in the cyclonic conditions of midwinter there may be sudden snowfalls in the south. The annual mean temperatures are 53.8° F (12.1° C) at Urbino, in the east, and 54.5° F (12.5° C) at Potenza, in Basilicata; the annual rainfall is, respectively, 35 inches and 39.6 inches. Along the Tyrrhenian coast and the Ligurian rivieras in the north, both temperature and rainfall are influenced by full exposure to the noonday sun, by the nearness of the sea, with its prevailing southwesterly winds, and by the Apennine range, which protects the area from the cold north winds. The eastern riviera has more rain than the western: rainfall at La Spezia, on the eastern riviera, is 45.2 inches, while at San Remo, on the western riviera, it is 26.7 inches. Farther south, where the coastal areas extend a great distance inland and are flatter, the mean temperature and annual rainfall are 58.6° F (14.8° C) and 30.3 inches at Florence and 61.9° F (16.6° C) and 31.4 inches at Naples. As a rule, the Tyrrhenian coast is warmer and wetter than the Adriatic coast. Both Calabria and Sicily are mountainous regions that are surrounded by the Mediterranean, and they therefore have higher temperatures than the high regions of the Italian mainland farther north. Winter rains are scarce in the interior and heavier in the west and north of Sicily. At Reggio di Calabria the annual mean temperature is 64.7° F (18.2° C) and rainfall is 23.5 inches; at Palermo, in Sicily, they are 64.4° F (18.0° C) and 38.2 inches. The sirocco, a hot, very humid, and oppressive wind, blows frequently from Africa and the Middle East. In Sardinia conditions are more turbulent on the western side, and the island suffers from the cold mistral blowing from the northwest and also from the sirocco blowing from the southwest. At Sassari, in the northwest, the annual mean temperature is 62.6° F (17° C) and the rainfall 22.8 inches, while at Orosei, on the east coast, the temperature is 63.5° F (17.5° C) and the rainfall 21.2 inches.

PLANT LIFE

The three vegetation zones

The native vegetation of Italy reflects the diversity of the prevailing physical environments in the country. There are at least three zones of differing vegetation: the Alps, the Po Valley, and the Mediterranean-Apennine area.

From the foot of the Alps to their highest peaks, three bands of vegetation can be distinguished. First, around the Lombard lakes, the most common trees are the evergreen cork oak, the European olive, the cypress, and the cherry laurel. Slightly higher, on the mountain plain, the beech is ubiquitous, giving place gradually to the deciduous larch and the Norway spruce. In the high-altitude zone, twisted shrubs, including rhododendron, green alder, and dwarf juniper, give way to pastureland that is covered with grasses and sedges and wildflowers such as gentian, dryad, rock jasmine, campion, sea bindweed, primrose, and saxifrage. Farther up there is curved sedge, with the dwarf willow and the lovely anthophytes. On the snow line there are innumerable mosses, lichens, and a few varieties of hardy pollinating plants, such as flags and saxifrage.

In the Po Valley almost nothing remains of the original forests; nearly all the vegetation has been planted or disposed by human activity. Poplars predominate where there is abundant water, but in the drier, more gravelly zones there are a few sedges. On the clayey upland plains, heather abounds, and there are forests of Scotch pine. There are the usual grasses beside the streams and in the bogs and water lilies and pondweed on the banks of the marshes. But the heavily predominant plants are the cultivated crops—wheat, corn (maize), potatoes, rice, and sugar beets.

In the Apennine zone along the whole peninsula, a typical tree is the holm oak, while the area closer to the sea is characterized by the olive, oleander, carob, mastic, and Aleppo pine. There is a notable development of pioneer sea grape on the coastal dunes. The Mediterranean foothill area is characterized by the cork oak and the Aleppo pine. Higher up, in southern Italy, there are still traces of the ancient mountain forest, with truffle oak, chestnut, flowering ash, Oriental oak, white poplar, and Oriental plane. There are quite extensive beech woods in Calabria (on La Sila and Aspromonte mountains) and Puglia, and the silver fir and various kinds of pine thrive in Abruzzi and Calabria. Where the forests have been destroyed in the strictly Mediterranean section of the Apennines, a scrub that is called maquis has grown up. On the island of Sardinia the destruction of the carob forests and on the Apulian Plain the decay of olive trees and shore vegetation have produced steppes of tough plants such as the various sorts of feather grass. Mountain meadowlands are found in Calabria and Basilicata, usually with vetch, bent grass, and white asphodel. The Apennine pasturelands are very much like those of the Alps. The papyrus is quite common in Sicily as a freshwater plant.

ANIMAL LIFE

The extent of animal life in Italy has been much reduced by the long presence of human beings. In the Alps there are quite a number of animals, such as marmots, that hibernate and others that change their protective colouring according to the season, such as the ermine, the mountain partridge, and the Alpine rabbit. Larger mammals include the ibex, which is protected in Gran Paradiso National Park, the chamois in the Central Alps, and the roe in the eastern Alps. The lynx, the stoat, and the brown bear (protected in Adamello and Brenta) are now rare. Alpine birds include the black grouse, the golden eagle, and, more rarely, the capercaillie, or wood grouse. Among the reptiles are vipers, and among the amphibians are the Alpine salamander and Alpine newt. Species that are found in the Alps also exist in other high mountain regions, where there are, however, more foxes and wolves. In Abruzzi the brown bear may be found, and on the island of Sardinia the fallow deer, the mouflon, and the wild boar are found. Among the freshwater fish are the brown trout, the sturgeon, and the eel. Among sea fish, besides common species such as the red mullet and the dentex, there are, especially in southern waters, the white shark, the bluefin tuna, and the swordfish. Among invertebrates, there is an abundance of red coral and commercial sponge on the rocks of the warm southern seas. In caves the greater horseshoe bat is found.

Remains of Roman tombs lining the Appian Way (begun 312 BC), Rome.
Fridmar Damm—De Wys Inc.

TRADITIONAL REGIONS

Italy is divided into 20 administrative regions, which correspond generally with historical traditional regions, though not always with exactly the same boundaries. A better-known and more general way of dividing Italy is into four parts: the north, the centre, the south, and the islands.

Problems of regional division

The north includes such traditional regions as Piedmont, which is characterized by some French influence and was the seat of united Italy's former royal dynasty; Liguria, extending southward around the Gulf of Genoa; Lombardy, which has long been noted for its productive agriculture and vigorously independent city communes and now for its industrial output; and the Veneto, once the territory of the far-flung Venetian empire, reaching from Brescia to Trieste in its greatest extent. The centre includes Emilia-Romagna, with its prosperous farms; the Marche, on the Adriatic side; Tuscany and Umbria, celebrated for their vestiges of Etruscan civilization and Renaissance traditions of art and culture; Latium, which contains the Campagna, whose beautiful hills encircle the eternal city of Rome; and the Abruzzi and Molise, regions of the highest central Apennines, which used to support a wild and remote people. The south (Mezzogiorno) includes Naples and its surrounding fertile Campania; the region of Puglia, with its great plain crossed by oleander-bordered roads leading to the low Murge Hills and the heel of Italy; and the poorer regions of Basilicata and Calabria. Finally, on the islands of Sicily and Sardinia are people who take pride in holding themselves apart from the inhabitants of mainland Italy. However, the south and the islands have changed a great deal since about 1960 and have become more modernized. Within these four main divisions, the variety of the much smaller traditional districts is very great and depends on history as well as on topography and economic conditions.

SETTLEMENT PATTERNS

Rural areas. The majority of the population of Italy live in cities and villages; only a fraction live in hamlets or in isolated houses.

In the long Alpine valleys the economy was always both agricultural and commercial, and there are many towns, such as Aosta and Bolzano, at the outlets of the lateral valleys. In settlements higher up or on the slopes of hills, an agricultural economy has remained predominant. On spurs of hillocks at the heads of valleys there are often old castles, originally built there for defense. The perpetual subdivision of landholdings makes a purely agricultural economy precarious in this region except in the upper Adige, where the Germanic system of primogeniture survived, producing the *masi*, family holdings that are passed on to the eldest son intact. Cattle raising remains profitable, but woodlands yield less return. These rural areas now also include an increasing number of skiing and tourist centres, such as Courmayeur and Cortina d'Ampezzo. Although these developments have reduced both the seasonal and the permanent migration away from this area, strictly rural life is, nevertheless, declining sharply here. In the band of Alpine and Apennine foothills, the villages, often situated on the knolls and flanks of the hills, are linked by roads that hold to the heights, away from the humid valley floors. Each village is usually grouped around a church, a castle, or a nobleman's palace, with its fields on the slopes around it and woodlands lower down. There are innumerable plum and cherry orchards and, above all, vineyards; their wines (Conegliano and Monferrato) are famous. Businesses are usually small or moderate in size. Lombardy is the only area in which the ancient rural way of life has been comprehensively displaced by the development of heavy industry. The population of its rural districts has been increased by migration from the neighbouring mountains and from the south. The Padano-Venetian-Emilian plain is the most important agricultural and stockbreeding region of Italy. The upland plain has now been virtually overrun by the great industrial centres such as Turin, Milan, and Busto Arsizio, but the lowland plain remains socially as well as economically rural. Wheat and corn are the most common crops, though each district has its specialty, such as sugar beet, grapes, and other fruit.

Apennine settlement patterns

Villages high in the Apennines are less prosperous than those of similar altitude in the Alps. They are still isolated, the ground is infertile, and land is rarely owned by those who work it. These hopeless conditions have caused the more enterprising residents to emigrate to the north, leaving the villages in an even more desperate situation. Tourism and the expansion of cottage craft industries, such as the porcelain making at Gubbio, near Perugia, have helped these towns survive. The lower hills and plains of Italy are covered with agricultural villages in which a wide variety of crops and vegetables are grown. The fields are heavily cultivated, but their yield is low; specialized cultivation is more profitable. In Puglia and Basilicata large farms are staffed by labourers who live in urban centres, such as Cerignola and Altamura, and travel to work in the countryside. Some fertile and well-watered plains, such as the Neapolitan countryside, have a high level of productivity, especially of market vegetables. Here there is direct ownership of land and fairly dense settlement. In Sicily, settlement is clustered in widely spaced, nucleated "peasant-cities." Wheat is extensively cultivated. Especially on the coasts, pastureland is extensive though not very profitable; there is efficient cultivation of grapes, olives, citrus fruits, and vegetables, all of which bring the island some revenue because they can be marketed as early produce. In Sardinia the settlement is sparse and mainly inland, because of the need, in historical times, to avoid the dangers of malaria in low-lying areas and also of the risk of attack by pirates. Although islanders, the Sardinians have never wanted to work on the sea, and most of their fishing industry is carried on by men from the mainland. There are extensive meadows and some forested areas survive.

Urban centres. From antiquity, Mediterranean peoples have had highly developed urban centres. For historical as well as geographic reasons, Italy has never been dominated by one city, each district tending to possess its own urban centre. Today, there are four cities with a population of more than 1,000,000 (Rome, Milan, Turin, and Naples); but many cities have a population of more than 100,000. Of these, almost half are on or near the sea; a similar proportion are in the north, and the rest are in the centre, the south, and on Sicily and Sardinia. This irregularity of urban settlement reflects the economic imbalance among different parts of the country. The distribution of Italian cities also reflects historical and geographic conditions. In the Po Valley, cities such as Milan, Pavia, and Cremona are well placed for commerce, being situated at the confluence of roads or rivers. Another group of cities are those on the coast, built at the mouths of rivers, or on lagoons protected by sandbars; these include Savona, Genoa, Naples, Messina, Palermo, Ancona, and Venice. A number of cities have, with increased population and industrialization, merged with neighbouring cities into enormous metropolitan complexes, sometimes characterized as mega-cities, such as that surrounding Milan. There are now several metropolitan areas in Italy, including Milan, Naples, Rome, Turin, Genoa, Florence, Palermo, and Bologna. (G.Na./R.L.K.)

The people

ETHNIC COMPOSITION

Italians cannot be typified by any one physical characteristic, a fact that may be explained by the past domination of parts of the peninsula by different peoples. The Etruscans in Tuscany and Umbria and the Greeks in the south preceded the Romans, who "Latinized" the whole country and maintained unity until the 5th century. Jews arrived in Italy during the Roman Republic, remaining until the present day. With the collapse of the Roman Empire in the West, Italy suffered invasions and colonization, which inevitably affected its ethnic composition. With some exceptions, the north was penetrated by Germanic tribes crossing the Alps, while the south was colonized by Mediterranean peoples arriving by sea. The Byzantines were dominant in the south for five centuries, coinciding with the supremacy of the Lombards (a Germanic tribe) in Benevento and other parts of the mainland. In the 9th century Sicily was invaded by the Saracens, who remained until the Norman invasion in the early 11th century. The Normans were succeeded by the Aragonese in 1282; in 1720 Sicily came under Austrian rule. This mixed ethnic heritage explains the smattering of light-eyed, blond Sicilians in a predominantly dark-eyed, dark-haired race. Except for the Saracen domination, the Kingdom of Naples, which formed the lower part of the peninsula, had a similar experience, whereas the northern part of Italy, separated from the south by the Papal States, was much more influenced by the dominant force of the Austrians. The Austrian admixture, combined with the earlier barbarian invasions, may account for the greater frequency of light-eyed, blond Italians orginating in the north. The ethnic mixing continues to the present day. Since the 1970s, Italy has been receiving immigrants from a number of Third World countries. A predominantly female migration from the Philippines and other Asian countries compares with a predominantly male influx from North Africa. Other African and Latin-American countries are also represented, as is eastern Europe with a more recent wave of immigrants. In total more than one million foreigners reside on Italian territory.

LINGUISTIC COMPOSITION

Standard Italian has only been in existence since the unification of Italy in the 1860s. The Italians were slow to adopt the parlance of the new nation-state, identifying much more strongly with their regional dialects. The eventual supremacy of the standard language owes much to the advent of television, which has introduced it into almost every home in the country. Emigration in the late 19th and early 20th century also played an important role in spreading the standard language; many local dialects had no written form, obliging Italians to learn Italian in order to write to their relatives. The extremely rich and, hitherto, resilient tapestry of dialects and foreign languages upon which standard Italian has gradually been superimposed reveal much about Italy's cultural history. Not surprisingly, the greatest divergence from standard Italian is found in border areas, in the mountains, and on the islands of Sicily and Sardinia. Only a few languages in limited geographic areas enjoy any legal protection or recognition. These are French, in Valle d'Aosta; German and Ladin in some parts of the Alto-Adige (Südtirol); Slovene in the province of Trieste; Friulian and Sardinian, spoken by the two largest linguistic minorities in Italy, received official recognition in 1992. Linguistic minorities persisting in the Alps are, broadly speaking, the result of migratory movements from neighbouring countries or changes in the borderline. The French and Franco-Provençal spoken in Valle d'Aosta date from union with Savoy, but the German spoken in the same area dates from the 12th-century emigration of German sheepherders from the upper valleys of the Rhône. The German spoken in the Alto-Adige dates back to Bavarian occupation in the 5th century, whereas that spoken in the provinces of Verona and Vicenza dates from a more recent colonization in the 12th century. Some of the Alpine areas have such a complex linguistic makeup that precise measurements of linguistic communities are impossible. In Friuli-Venezia Giulia, for example, many communes are bi-, tri-, and even quadrilingual, as in the case of Canale, where Slovene, Italian, German, and Friulian coexist. In certain Occitan-speaking parts of Piedmont, Italian is the official language, Occitan is spoken at home, and the Piedmontese dialect is used in trading relations with people from lowland areas. Farther south, in Abruzzi, Basilicata, Calabria, Puglia, and Sicily, isolated linguistic communities persist against the odds. A dialect of Albanian known as Arbëresh is spoken by the descendants of 15th-century Albanian mercenaries; Croatian, the smallest minority language, spoken by some 2,000 people, has survived in splendid isolation in Campobasso province in Molise; and Greek, or "grico" (of uncertain origin), may be heard in two areas in Calabria and Puglia. Catalan, too, has survived in the town of Alghero in the northwest of Sardinia, dating from the island's capture by the crown of Aragon in 1354.

Legal recognition of minority languages

RELIGION

Roman Catholicism became the official religion of the Italian state in 1929 with the signing of the Lateran Treaty between the fascist leader Benito Mussolini and Pope Pius XI. More than 90 percent of the population declare themselves Roman Catholics, although the number of practicing Catholics is declining. A concordat signed in 1984 ended the church's position as the state religion, abolished compulsory religious teaching in public schools, and reduced state financial contributions to the church. An estimated 225,000 people worship in the Protestant church, including Lutherans, Methodists, Baptists, and Waldensians. They are all members of the Federation of Evangelical Churches in Italy (Federazione delle Chiese Evangeliche in Italia) founded in 1967. The Jewish community declined in numbers during World War II as a result of Nazi persecution. Since then its numbers have increased mainly through immigration. In 1987 the Jewish community obtained special rights from the Italian state allowing them to abstain from work on the Sabbath and to observe Jewish holidays. Albanian communities in two dioceses and one abbey in the Mezzogiorno practice the Eastern Orthodox rite. The most recent wave of migration from the Third World has brought with it people of non-Christian religious beliefs, the most significant minority being Muslim.

DEMOGRAPHIC TRENDS

Throughout the centuries, Italy's population curve has undergone many changes, often in parallel development with population trends in other European countries. The mid-14th-century plague reduced the peninsula's population considerably, and a long period of population growth ended at the beginning of the 17th century. From the early 18th century until unification in the 1860s, a slight, steady growth prevailed, although it was interrupted during the Napoleonic Wars. From the latter half of the 19th century to the latter half of the 20th century, the population more than doubled, despite high levels of emigration. An interesting feature to note is that natural population increase was frequently highest during the decades of highest emigration, although there is no obvious causal relationship between the two.

Italy's overall demographic trends are still fairly consistent with those of other advanced western European countries. A fall in the fertility and mortality rates, however, manifested itself more than a decade later in Italy than in those countries because of Italy's slower socioeconomic development before World War II. The growth rate of the population is gradually slowing, with most of the increase coming from immigration; birth rates and death rates are virtually identical. However, the national figures conceal contrasting regional trends. In general, the birth rate and average family size are higher in the south of Italy than in the north, although populations in Molise, Basilicata, and Calabria are declining through continued emigration. The mortality rate is slightly lower in the south than in the north as a result of improved medical care and a younger population; in certain northern regions, especially Liguria, populations are beginning to decrease because the birth rate is falling faster than the mortality rate. As regards the country as a whole, life expectancy has risen during the second half of the 20th century, reflecting higher nutritional, sanitary, and medical standards. The majority of the population are between 20 and 69 years old, with decreasing numbers below 10 and increasing numbers over 75, especially among women.

Regional trends

Internal migration patterns. Since the unification of Italy, internal movements have followed a regular pattern—south to north and east to west. People have moved from the southern regions and Sicily to the central regions of Latium and Tuscany and to the northwest—to Lombardy, Liguria, and Piedmont. They moved in the same way from Veneto to the northwest. Movement from Emilia-Romagna, Marche, and Umbria to regions in the northwest has also been significant. Population movement was relatively slight during the fascist era between the wars, when permits were required for movement inside the country. Exceptionally, substantial numbers of Italians seeking work at the huge Lingotto factory run by Fiat were granted permits to go to Turin.

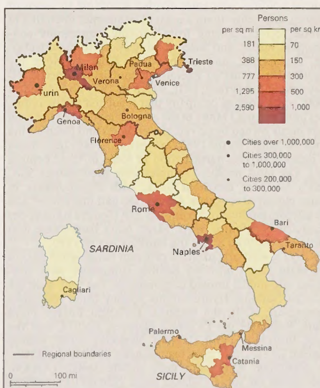
After World War II and the demise of fascism, Italy entered a period of unprecedented economic growth and high population mobility. The prosperity of the urban areas, especially the industrial triangle of Lombardy-Piedmont-Liguria, contrasted with continuing hardship and poverty in the upland and rural areas, especially in the south. Rapid industrialization in the urban centres acted as a strong "pull" factor, encouraging rural workers to abandon the land and head for the cities. The disparity between wealth and employment in urban and rural areas triggered a period of intense rural depopulation from the uplands in the Alps, the Apennines, Sicily, and Calabria, and an influx of migrants to Rome, Milan, Turin, and Genoa. This movement continues today, although the slowing of economic growth has reduced the "pull" factor exerted by the industrial areas to more of a myth propagated by television; unemployment runs high, especially among the young, and those moving away from many rural districts are now exceeded by those returning.

Emigration and immigration. Between 1876 and 1970 an estimated 25 million Italians left the country in search of work. Of these, 12 million left for destinations outside Europe. In the 1860s, transatlantic migration was most frequent among northern Italians and was often associated with certain trades; for example, farmers, artists, and street traders tended to emigrate to America. Two decades later, however, the trend had become a mass phenomenon, with the main protagonists increasingly emanating from the south. Their principal destination was the United States, favoured by more than half the emigrants, with the others choosing Argentina, Brazil, and Canada. Some also went to Australia. In the 1920s the United States introduced strict immigration laws, and economic conditions in Brazil and Argentina deteriorated to such an extent that transatlantic emigration was stymied. In addition, the fascist regime opposed emigration, and during World War II emigration halted almost completely. After 1945 destinations were mainly European, the most popular being France initially and then West Germany and Switzerland. The nature of emigration patterns changed, becoming less stable. In many cases the emigrants were mostly male, as some European countries refused entry to workers' relatives because of housing shortages. Often Italian workers would remain abroad for short periods of time, returning every so often to Italy. On the eve of the 1973 oil embargo more than 850,000 Italians were working in Switzerland and countries of the European Community (EC), but the ensuing recession caused Switzerland to implement forced repatriation policies, and many Italians were "squeezed" back home by unemployment in EC countries.

Movement to EC countries

In 1972 Italy for the first time registered more people entering the country than leaving. A simultaneous reduction of immigrant worker figures in West Germany and Switzerland suggested that significant numbers of Italians were returning home, but they did not account for all the entrants. Increasing numbers of migrants from Asia, Africa, and Latin America were arriving in Italy. For several years the scale of the influx of Third World immigrants was difficult to assess, as no policy existed either to measure or to control the phenomenon. It was not until 1986 that the first policy measures were introduced. By 1989 the numbers and distribution of foreigners became clearer: of 1,100,000 foreigners, 960,000 were from outside the EC, and nearly half were concentrated in the big urban centres of Rome, Turin, Genoa, Florence, and Palermo. The majority of new arrivals settled in the north and centre of Italy, but the south had a relatively higher proportion of African immigrants than the north.

The collapse of communist regimes in eastern Europe brought fresh waves of immigrants from Poland, Romania, Albania, and the Yugoslav region. Many arrived via



Population density of Italy.

seaports on the Adriatic coast, claiming refugee status. Some were repatriated, but others were relocated to inland destinations.

The economy

AN OVERVIEW

The Italian economy has progressed from being one of the weakest economies in Europe following World War II to being one of the most powerful. Its strengths are its metallurgical and engineering industries; its weaknesses are a lack of raw materials and energy sources. More than four-fifths of Italy's energy requirements are imported. Nonetheless, the chemical and textile industries are among Italy's largest. A strong entrepreneurial bias, combined with liberal trade policies following the war, enabled manufacturing exports to expand at a phenomenal rate, but a cumbersome bureaucracy and insufficient planning hindered an even economic development throughout the country. Services, particularly tourism, are also very important. At the end of the 20th century, Italy, seeking balance with other EU nations, brought its high inflation under control and adopted more conservative fiscal policies, including sweeping privatization.

Although the Italian economy was a relative latecomer to the industrialization process, business in the north of the country has caught up and overtaken many of Italy's western European neighbours. Southern Italy, however, has lagged behind. The percentage of the labour force working in agriculture is often taken as an indication of the rate of industrialization and wealth of a nation, and in Italy's case the figures illustrate an imbalance between north and south. In 1990, 5.3 percent of the Italian population worked on the land, compared with an EU average of 7 percent. Yet, there were as many agricultural labourers from the south, in just 8 regions, as they were from the 12 regions in the north and centre. Calabria and Basilicata have the largest concentrations of farm labourers.

Although Italy is not self-sufficient agriculturally, certain commodities form an important part of the export market. Notably, the country is a world leader in olive oil production and a major exporter of rice, tomatoes, and wine. Cattle raising, however, is less advanced; most meat and dairy products are imported.

Public and private sectors. The Italian economy is mixed, and until the beginning of the 1990s the state owned a substantial number of enterprises. At that time the major state holding corporations were organized as pyramids, with a holding company at the top, a middle layer of financial holding companies divided according to sector of activity, and below them a mass of companies operating in diverse sectors, ranging from banking, expressway construction, media, and telecommunications to manufacturing, engineering, and shipbuilding. (One example, the Institute for Industrial Reconstruction [Istituto per la Ricostruzione Industriale; IRI], set up in 1933 and closed in 2000, was a holding company that regulated public industries and banking.) Many of the companies on the base level were partly owned by private shareholders and listed on the stock exchange. By the 1980s moves had already been made to increase private participation in some of these companies. The most notable examples were Mediobanca SpA, Italy's foremost merchant bank, with shareholdings in major industrial concerns; Alitalia, the national airline; and the telecommunications company SIP (Società Italiana per l'Esercizio Telefonico SpA). Many other banks were also partially privatized under the Banking Act of 1990.

Privatization program

In 1992 a wide privatization program began when four of the main state-controlled holding companies were converted into public limited corporations. The four were the IRI, the National Hydrocarbons Agency (Ente Nazionale Idrocarburi; ENI), the National Electrical Energy Fund (Ente Nazionale per l'Energia Elettrica; ENEL), and the State Insurance Fund (Istituto Nazionale delle Assicurazioni; INA). Other principal agencies include the Azienda Nazionale Autonoma delle Strade Statali (ANAS), responsible for some 190,000 miles (350,000 km) of the road network, and the Ente Ferrovie dello Stato (FS; "State Railways"), controlling the rail network.

The private sector is characterized by a multitude of small companies, many of which are family-run and employ few workers outside the family. In fact, in 1999, firms with fewer than 20 employees represented 98 percent of total firms, reflecting a decade-long trend that showed a decline in large production units and an increase of smaller, more specialized ones. This trend was especially pronounced in the automobile industry, textiles, electrical goods, and agricultural, industrial, and office equipment.

Following World War II, the economy in the south was mainly dominated by the interests of the government and the public sector. The Southern Development Fund (Cassa per il Mezzogiorno), a state-financed fund set up to stimulate economic and industrial development between 1950 and 1984, met with limited success. It supported early land reform—including land reclamation, irrigation work, infrastructure building, and provision of electricity and water to rural areas—but did little to stimulate the economy. Later the fund financed development of heavy industry in selected areas, hoping that major industrial concerns might attract satellite industries and lay the foundation for sustained economic activity. Yet these projects became known as "cathedrals in the desert"; not only did they fail to attract other smaller industries, they also suffered from high absenteeism among workers. The most successful project was undertaken by Finsider, which in 1964 opened Europe's most modern steelworks, in Taranto.

Postwar economic development. The development of the Italian economy after World War II was one of the country's major success stories. Economic reconstruction was followed by unprecedented economic growth between 1950 and 1963. Gross domestic product (GDP) rose by an average of 5.9 percent annually during this time, reaching a peak of 8.3 percent in 1961. The years from 1958 to 1963 were known as Italy's economic miracle. The growth in industrial output peaked at over 10 percent per year during this period, a rate surpassed only by Japan and West Germany. The country enjoyed practically full employment, and in 1963 investment reached 27 percent of GDP. The success was partially due to the decision to foster free market policies and to open up international trade. Italy was an enthusiastic member of the EC (European Community, later part of the European Union [EU]) from the very beginning, and its membership favoured the Italian manufacturing industry, which expanded enormously during this period. Certain products, such as Olivetti typewriters and Fiat automobiles, dominated European and world markets in just a few years. The economy slowed down after 1963 and took a downturn after the 1973 increase in world petroleum prices, but again prospered by the late 1980s.

Italy's economic miracle

Recent economic trends. The economy entered the mid-1980s with a healthy growth rate, which it maintained through the end of the century. However, there were serious battles to be waged: against inflation, a trade deficit, currency restrictions, and tax evasion.

Inflation reached nearly 22 percent in 1980. This was principally due to union strength in wage bargaining throughout the 1970s and a mechanism called the *scala mobile,* which adjusted wages to inflation on a quarterly basis for all wage and salary earners. The high degree of job security enjoyed by the Italian workforce raised production costs, which in turn contributed to inflation. Beginning with a decree in 1984 that imposed a ceiling on payments, the *scala mobile* was gradually dismantled (and abolished in 1992) under pressure from the employers' association, the Confederation of Industries (Confindustria). These actions were reflected in a sharp fall in inflation to 12 percent in 1984 and down to 4.2 percent in 1986. However, a three-year contract signed in 1987 between Confindustria and trade unions representing all civil servants and some private industrial workers awarded pay raises over the rate of inflation, and by 1991 inflation was again up to 7 percent—3 percent higher than in Germany or France. In 2000 inflation in Italy was at 10 percent. Overall, however, the inflation rate was three times less throughout the 1990s than in the 1980s.

The *scala mobile*

Italy's public debt grew steadily throughout the 1980s despite a series of emergency measures designed to reduce public borrowing. By 1991 public debt exceeded GDP, and the cost of servicing it was more than $100 billion, accounting for the entire government budget deficit for the year. At the turn of the 21st century, Italy's public debt still exceeded its GDP.

Italy underwent currency reform in the 1980s and '90s in an effort to come in line with the fiscal standards set by the EU. At the end of the century, Italy joined the single currency of the EU, adopting the euro in 1999.

In recent years the Italian economy has been dogged by the government's inefficient levying of direct taxes. Since the creation of the republic after World War II, the economy has relied on public loans to finance public works and enterprises, and many Italians did not start paying income tax until the 1970s. Italy also has a thriving underground economy that inevitably deprives the state of revenue. While indirect taxes, including VAT (value-added taxes), have been raised several times throughout the 1980s, moves to enforce payment of direct taxes have met with resistance. In 1985 a bill was introduced to curtail tax evasion among the self-employed, leading to a one-day national strike. The 1990 budget also included measures to reduce tax evasion. The names of the country's top taxpayers are publicized annually in an attempt to encourage compliance with tax laws.

During the 1990s the annual GDP growth rates were very modest. In 2000, in response to a healthy international economy and to steps taken to improve the Italian finance system—including reduced public spending and increased taxation—the GDP grew 3 percent, its biggest increase since 1988, but it was unclear whether this recovery would be sustained.

RESOURCES

The Italian peninsula is a geologically young land formation and therefore contains few mineral resources, especially metalliferous ones. What few exist are poor in quality, scant in quantity, and widely dispersed. The meagreness of its natural resources partially explains Italy's slow transition from an agricultural to an industrial economy, which began only in the late 19th century. The lack of iron ore and coal especially hindered industrial progress, impeding the production of steel necessary for building machines, railways, and other essential elements of an industrial infrastructure.

Paucity of natural resources

Iron and coal. Half of Italy's iron output comes from the island of Elba, one of the oldest geologic areas. Another important area of production is Cogne in the Alpine region of Valle d'Aosta; this deposit lies at 2,000 feet (610 metres) above sea level. Little iron-bearing ore has been produced in Italy since 1984. Coal is found in small amounts principally in Tuscany, but it is of inferior quality, and its exploitation has been almost negligible. The vast majority of Italy's coal is imported, mostly from Russia, South Africa, the United States, and China.

Mineral production. In recent decades, production of almost all Italy's minerals steadily decreased, with the exception of rock salt, petroleum, and natural gas. In the early 1970s Italy was a major producer of pyrites (from the Tuscan Maremma), asbestos (from the Balangero mines near Turin), fluorite (fluorspar; found in Sicily and northern Italy), and salt. At the same time, it was self-sufficient in aluminum (from Gargano in Puglia), sulfur (from Sicily), lead, and zinc (from Sardinia). By the beginning of the 1990s, however, it had lost all its world-ranking positions and was no longer self-sufficient in these resources. Fuel deposits, too, were unable to keep pace with the spiraling demands of energy-hungry industries and domestic consumers. Although domestic production figures rose throughout the late 20th century, Italy remains a net energy importer. Small amounts of oil and natural gas used to be produced in the Po Valley in the 1930s, and asphalt was produced in Ragusa in Sicily. This exploitation was followed by further oil discoveries in the Abruzzo and richer amounts again in Ragusa and in nearby Gela. Natural gas is the most important natural resource in the peninsula, found mainly on the northern plain but also in Basilicata, Sicily, and Puglia.

Italy is the world's leading producer of pumice and feldspar. Another mineral resource for which Italy is well-known is marble, especially the world-famous white marble from the Carrara and Massa quarries in Tuscany. However, the reputation of these exceptional stones is disproportionately large when compared with the percentage of GNP accounted for by their exploitation.

AGRICULTURE, FORESTRY, AND FISH PRODUCTION

Like other branches of the Italian economy, agriculture has been characterized historically by a series of inequalities, both regional and social. Until the Land Reform Acts of 1950, much of Italy's cultivable land was owned and idly managed by a few leisured noblemen, while the majority of agricultural workers struggled under harsh conditions as wage labourers or owned derisory plots of land, too small for self-sufficiency. Agricultural workers had few rights, and unemployment ran high, especially in Calabria, where the impetus for land reform was generated. Reform entailed the redistribution of large tracts of land among the landless peasantry, thereby absorbing greater amounts of labour and encouraging more efficient land use.

Although partially successful, the reform created many farms that were still too small to be viable and plots that were scattered in parcels and often located in unfertile uplands. Another negative aspect of the reform was that it had the effect of damaging the social structure of rural communities. Initially, the European Economic Community (EEC; later the EC) did little to help Italy's small farmers, located primarily in the south, while wealthier, larger farms in the north benefited from EEC subsidies. However, in 1975 specific aid was directed at upland farmers, and in 1978 another package provided them advisory support and aid for irrigation. Today most farms are owned and operated by families.

Since World War II, Italy has maintained a negative trade balance in agricultural products, many of which are consumed domestically because of the country's high population density. The majority of foreign agricultural and food-related trade is with other EU countries, in particular with France and Germany.

Italy's plains constitute only one-fourth of the land under cultivation, indicating widespread cultivation of hilly environments where agriculture has been possible only as a result of modifying the natural landscape and resources through terracing, irrigation, and soil management. The most fertile area is the Po Valley, where precipitation is fairly evenly distributed throughout the year, but mean rainfall decreases southward. Coastal areas in Puglia, Sicily, and Sardinia may register only 12–16 inches (300–400 mm) of annual precipitation, compared with 118 inches in Alpine regions.

In general, agricultural land use is divided into four types—field crops, tree crops, pasture, and forestry.

Agricultural land use

Field crops. While prime minister in 1922–43, Benito Mussolini strove to make Italy self-sufficient in the production of wheat, but since that time the land given over to its cultivation has been reduced from more than 12 million acres (4.8 million hectares) to just over 5 million acres. Hard wheat used for making pasta is traditionally grown in the south, whereas soft wheat used for making bread, biscuits, and pizza crust predominates in the northern lowlands. Yields in the north can be up to three times those in the south because of improved mechanization and more suitable terrain.

Italy is a major exporter of rice, which is grown mostly on the Po plain. Corn (maize) also is grown in this area. Of the other field crops, tomatoes are the most important for domestic and export markets. Naples and Emilia-Romagna specialize in this crop. In the late 20th century, the area given over to growing tomatoes increased more than twofold, and production quadrupled as a result of improved production techniques.

Olives and grapes

Tree crops. Olives and grapes are Italy's two most lucrative agricultural exports. Olive production is suited to the arid conditions of Puglia, Sicily, and Calabria, the oil content being enhanced by the long, dry summers. The output is erratic, however, as the olives are susceptible to late frosts. Italy is the world's biggest exporter of olive oil, although Spain dominates the more lucrative sector of table olives. While olives are traditionally grown in conjunction with other crops or livestock raising, nearly half the olive-producing land now excludes other types of cultivation, reflecting the demise of traditional peasant farming methods.

Wine is produced in every region of Italy and, together with olive oil, enjoys a positive trade balance. Competition is stiffening, however, with the burgeoning eastern European market undercutting western European prices. Much of the heavier wine from the south is used to produce vermouth or marsala, while the best-known wines—Soave, Valpolicella, Barolo, and Asti—are produced in the north.

About three-fifths of Italy's citrus fruit production is Sicilian, with most of the rest growing in sheltered and irrigated lowlands in Calabria and Campania. Deciduous fruits, on the other hand, are widespread. Campania is best known for its cherries, apricots, nectarines, and hazelnuts, while Emilia-Romagna produces mostly peaches, plums, and pears. Sicily and Puglia produce almonds.

Pasture. Pastureland makes up about one-sixth of land use. Meat production in Italy is traditionally weak. Cattle production was relatively stagnant in the second half of the 20th century. There is a marked geographic difference in the distribution of farms; while bovine, swine, and aviculturist farms are mainly found in the north, ovine farms are more widespread in the south. Butter production satisfies domestic consumption, and many cheeses, including Gorgonzola and Parmesan, are made for export. Raising buffalo is a popular activity in Tuscany and Campania, where their milk is used for mozzarella cheese. The production of goats' milk is still modest, although it has become more lucrative; goats' milk is regarded as a luxury item for the urban market instead of peasant fare. The breeding of pigs has increased most dramatically, mostly in the northern regions of Lombardy and Emilia-Romagna; the air-dried hams (prosciutto) of Parma are world-famous, as are salami, bologna (known as mortadella in Italy), and other cured products. Peasant families traditionally keep pigs for their own consumption. Competition from other EU countries has threatened the Italian meat market, which suffers from high production costs because of the necessity for irrigation.

Forestry. Italian forestry has suffered from overexploitation in the past, first in antiquity by the Romans and then again in the 19th century, when much wood was needed for building mine shafts and railway sleepers. Less than one-fourth of the land is classified as forest and other woodland. Strenuous efforts to reforest certain areas are gradually producing positive results; for example, at the end of the 20th century, the production of roundwood, after dipping by 40 percent in the mid-1970s, nearly returned to the high levels it had maintained in the 1960s.

Broad-leaved forests

Most of Italy's forest area is made up of broad-leaved trees, with conifers making up about one-fifth of the total. Broad-leaved forests are fairly well spread over the country, with the exceptions of Puglia, Sicily, and Sardinia. Conifers are for the most part concentrated in the Alpine foothills, especially in the Alto Adige adjacent to the Austrian border. Chestnut forests are widespread in the northern Apennines and the Calabrian Sila. The North Italian Plain, Puglia, and the southern half of Sicily are virtually devoid of woodland.

Fishing. Italian fish production doubled in the last four decades of the 20th century, but consumption is met mainly by imports. About four-fifths of the fish come from the Mediterranean and the Black Sea and about one-tenth from the Atlantic Ocean, the remaining one-tenth coming from inland waters.

INDUSTRY

Mining and quarrying. Mining is not an important sector of the Italian economy. Minerals are widely dispersed, and, unlike other industries, mining and quarrying traditionally have been more prevalent in the south than in the north. In the late 20th century, increased demand for construction materials and fertilizers led to the expansion of northern-based quarrying industries specializing in lime and chalk (for the production of fertilizers and cement, an important industry), along with coloured granites and marbles. Conversely, in the north, extraction of metalliferous minerals (such as iron ore, manganese, and zinc) has declined. Nonmetalliferous minerals, including graphite, amianthus (a type of asbestos), and coal, shared a similar fate throughout Italy. As a primary fuel, coal satisfied less than 6 percent of the country's energy demands at the end of the 20th century. Mining has fared badly on the islands, where it once prospered, with a decline in the extraction of sulfur in Sicily and of lead and zinc in Sardinia. Industrial minerals that remain significant are barite, cement, clay, fluorspar, marble, talc, feldspar, and pumice.

Manufacturing. The most remarkable feature of Italian economic development after World War II was the spectacular increase in manufacturing and, in particular, manufacturing exports. The most significant contributory factors to this growth were the Marshall Plan (1948–51), a U.S.-sponsored program to regenerate the postwar economies of western Europe; the 1952 foundation of the European Coal and Steel Community (ECSC), later under the European Federation of Iron and Steel Industries; the start in 1958 of the EEC, which contributed to the liberalization of trade; and the abundance of manpower that fueled the growth of northern industrial concerns.

Postwar manufacturing

The material that transformed the Italian economy with a flourish was steel. Despite the lack of mineral resources, the Italian government opted to join the ECSC at its inception, and skeptics watched as Italian steel developed so quickly that by 1980 it accounted for more than one-fifth of production in the EC (which by then had nine members) and in western Europe. Moreover, Italy was second only to West Germany among western European steel producers. Steel formed the backbone of the metallurgical and engineering industries, known as *metalmeccanica.* These enjoyed their heyday between 1951 and 1975, when mechanical exports rose 20-fold and the workforce employed in the industries doubled. The number of people working in the automobile industry tripled, and metallurgical exports increased 25 times. The steel industry, which declined in the last decades of the century, was privatized in 1992–97.

The main branches of *metalmeccanica* included arms manufacture, textile machinery, machine tools, automobiles and other transport vehicles, and domestic appliances. The automobile industry has been dominated by Fiat since the founding of the company in Turin in 1899. Milan, home of Alfa Romeo and Lancia, and Brescia became the other main automaking centres until Alfa Romeo opened its plant near Naples, leading to a decentralization of the industry. Automobile production took off in the 1950s and soared until the mid-1970s, when it began to stagnate. In the 1980s imports from Japan and an economic recession further dampened the industry, though new markets were opened in eastern Europe at the end of the Cold War in the early 1990s. Today Italy has one of the highest numbers of cars per capita in the world.

Notable large firms notwithstanding, the manufacturing sector is characterized by the presence of small and medium-size industries, which are found mainly in northeastern and north-central Italy. This area, concentrated in industrial districts within Veneto, Emilia-Romagna, and Tuscany, is referred to as the "third Italy" to distinguish it from the "first Italy," represented by the industrial triangle formed by the cities of Milan, Turin, and Genoa, and from the "second Italy," which includes the Mezzogiorno. Each industrial district in the third Italy generally specializes in a particular area of light manufacturing, such as textiles or paper products, although more traditional manufacturing is also present. For instance, in Prato, Tuscany, the specialty is textile products; Sassuolo and Cento, both in Emilia-Romagna, engage in ceramic tile production and mechanical engineering, respectively; while Nogara, in Veneto, is known for wooden furniture.

Italy dominated the postwar domestic appliance market, which boomed until the first international oil crisis, in 1973, when small businesses were hard-hit by the increase in energy prices. Olivetti and Zanussi were market leaders, and Italian-produced "white goods," such as refrigerators and washing machines, were much in demand. The textile industry has been important in Italy since the Middle Ages, when Lombardy and Tuscany were leading centres for the wool and silk industry. Other important products now include artificial and synthetic fibres, cotton, and jute yarn. Textiles and leather goods were surpassed by the metallurgical sector in the 1960s, but they remain important components of manufacturing.

The chemicals industry is one of the more recent members of the Italian industrial family. It is often categorized into primary chemicals, dominated by giant enterprises, including Montedison, ENI Chem, and SNIA; secondary chemicals, made up of thousands of firms; and a third component comprising firms financed by foreign capital. From 1868 until World War I the chemical industry was restricted to products such as fertilizers and fungicides, but the oil discoveries of the 1950s opened up the vast field of petrochemicals. The discovery, too, of natural gas near Ravenna triggered the production of synthetic rubber, resins, artificial fibre, and more fertilizers. The industry boomed until the 1973 oil crisis launched a protracted slump, which rebounded somewhat in the 1980s as the sector was rationalized. During the 1990s the chemical industry was confronted with limits defined by environmental protection policies and the restructuring of small and medium-size enterprises.

The food and beverage industry is also important, in particular the traditional products olive oil, wine, fruit, and tomatoes. Additionally, the pulp and paper, printing and publishing, and pottery, glass, and ceramics industries are prominent.

Construction. The housing sector was affected by three main factors following World War II: the postwar economic boom, massive rural-to-urban migration, and government incentives to the private construction sector. Approximately 500,000 homes were destroyed in the war and another 250,000 severely damaged. A period of frenzied building ensued, reaching a peak during 1961–65, when an average of 380,000 houses were built each year. Much of the building was undertaken by private companies that engaged heavily in speculative construction and paid scant regard to regulations. This led to overcrowding and a severe lack of services in peripheral urban areas. The problem was exacerbated by the migration of hundreds of thousands of southern Italians to the big northern towns in search of work.

Construction slackened during the late 1960s and '70s as a result of economic recession, although many Italians were still living in substandard dwellings and awaiting rehousing. A 1980 earthquake in the Naples area destroyed a quarter of a million homes and resulted in a localized building boom lasting almost a decade. In the late 1990s the construction sector showed signs of recovery mainly related to investments in public works and the availability of financial incentives for residential housing.

Energy. Italy's lack of energy resources undoubtedly hindered the process of industrialization on the peninsula, but the limited stocks of coal, oil, and natural gas led to innovation in the development of new energy sources.

Pioneering of hydroelectricity

It was the dearth of coal in the late 19th century that encouraged the pioneering of hydroelectricity, and in 1885 Italy became one of the first countries to transmit hydroelectricity to a large urban centre—from Tivoli to Rome, along a 5,000-volt line. Rapid expansion of the sector developed in the Alps (with water passing efficiently over nonporous rocks) and also in the Apennines (with less efficient transport over porous rocks). Though uneven precipitation on the peninsula marred continuing growth in hydroelectricity, it comprised a healthy slice of the country's energy consumption by 1920. In the aftermath of World War II, more than half of Italy's electric power was accounted for by hydroelectricity, but there was little room left for expansion, and the country was in need of energy to feed its rapid industrialization.

This led to the development of thermal electricity generation fired by coal, natural gas, oil, nuclear power, and geothermal energy. In 1949 oil was discovered off Sicily, but supplies were limited, and Italy began to rely heavily on imported oil, mainly from North Africa and the Middle East. With oil in such short supply, Italy was, not surprisingly, at the vanguard of nuclear research, and by 1965 three nuclear power stations were operating on Italian soil; a fourth opened in 1981. Nonetheless, by 1987, nuclear power accounted for only 0.1 percent of Italy's total electricity production, and a public referendum of the same year led to the decommissioning of all four plants.

Natural gas has been the most significant discovery. It was first found in the 1920s, and its most important exploitation was in the Po Valley. Later exploration focused on offshore supplies along the Adriatic coast. Increased reliance on imports began in the 1970s, and by the beginning of the 21st century about three-fourths of Italy's natural gas was imported, primarily from Algeria, Russia, and The Netherlands. There are about 19,000 miles (30,000 km) of pipelines. The use of natural gas has risen at the expense of oil, which is currently the dominant form of energy in Italy, comprising about half of total consumption.

One-fifth of Italy's power comes from hydroelectricity. With demand greater than supply, Italy imports electricity from its neighbours, primarily Switzerland and France.

Service industries and tourism. The service sector is one of the most important in Italy in terms of the number of people employed. If the definition extends to cover tourism, the hotel industry, restaurants, the service trades, transport and communications, domestic workers, financial services, and public administration, well over half of the workforce operates in the sector. A fully accurate measure is impossible, however, because of the existence of a burgeoning black market.

A plentiful supply of labour has nourished the service sector, especially in the large urban areas, since the 1950s. This labour came initially from rural areas of northern Italy such as Veneto; later the Italian peasantry from the Mezzogiorno migrated north; and more recently immigrants from less-developed countries—many of whom work for low wages, without job security, and under substandard work conditions—have filled low-grade urban service jobs. High-level service jobs include those involved with information technologies, which are used by one-third of Italian business. Factors that have contributed to the growth of the service sector include the rise in the standard of living in Italy and Europe in general, leading to an increase in mobility, financial transactions, business, demand for leisure activities, and tourism.

Tourist destination

Italy is renowned as a tourist destination; it attracted more than 30 million foreign visitors annually at the end of the 20th century. Conversely, less than one-fifth of Italians take their holidays abroad. The tourist industry in Italy experienced a decline from 1987 onward, including a slump during the First Persian Gulf War and the world recession, but it rebounded later in the 1990s, posting gains in the number of overseas and domestic tourists. In addition, the Jubilee celebrations promoted by the Roman Catholic church in 2000 to mark the advent of its third millennium attracted millions of tourists to Rome and its enclave, Vatican City, seat of the church.

The tourist industry has flourished under both national and international patronage. The most popular locations, apart from the great cultural centres of Rome, Florence, Venice, and Naples, are the coastal resorts and islands or the Alpine hills and lakes of the north; the Ligurian and Amalfi rivieras; the northern Adriatic coast; the small islands in the Tyrrhenian Sea (Elba, Capri, and Ischia); the Emerald Coast of Sardinia; Sicily; Gran Paradiso National Park and the Dolomites; and Abruzzo National Park.

Labour and taxation. Women comprise less than two-fifths of the labour force, though they are more likely to take on fixed-term and part-time employment than men. The activity rate of male employment is consistent throughout Italy, but females have a much lower rate of participation in the south.

Scala mobile

Due to the *scala mobile,* which adjusted wages to inflation, Italian workers benefited from high job security for decades after World War II. Beginning in the 1980s, though, as the government moved to get inflation under control, the *scala mobile* came under attack.

The strength of trade unions was in decline by the end of the 20th century, but large general strikes were not uncommon. The right to strike is guaranteed by the constitution and remains a very potent weapon in the hands of the trade unions. Three major labour federations exist, each closely tied to different political factions: the General Italian Confederation of Labour (Confederazione Generale Italiana del Lavoro; CGIL), which is tied to the left; the Italian Confederation of Workers' Unions (Confederazione Italiana di Sindicati Liberi; CISL), with ties to the Catholic movement; and the Italian Labour Union (Unione Italiana del Lavoro; UIL), related to the secular parties. A number of independent unions are also active, especially in the public service sector. They increasingly challenge the monopoly of the three confederations on national contractual negotiations and are quite militant.

The government has undertaken reforms in tax collection. Historically, it has been unsuccessful in gathering income taxes with consistency, in part because of tax evasion and a black market on goods.

FINANCE

Italy's financial and banking system has a number of unique features, although its framework is similar to that of other European countries. The Bank of Italy is the central bank and the sole bank of issue. Monetary policy is vested in the Interministerial Committee for Credit and Savings, headed by the minister for the economy and finance. In practice, the Bank of Italy enjoys wide discretionary powers and plays an important role in economic policy making. In addition to the execution of monetary policy, its primary functions include the control of credit and the formation and execution of monetary policy.

There are three main types of banking and credit institutions. First, there are the commercial banks, which include three national banks, several chartered banks, popular cooperative banks—whose activities do not extend beyond the provincial level—and ordinary private banks. Second, there is a special category of savings banks organized on a provincial or regional basis. Finally, there are the investment institutions, which collect medium- and long-term funds by issuing bonds and supply medium- and long-term credit for industry, public works, and agriculture. The 1990 Banking Act reduced the level of public ownership of banks and facilitated the raising of external capital. All remaining controls on capital movements were also lifted, enabling Italians to bank unlimited amounts of foreign capital in Italy.

There are many institutes of various kinds supplying medium- and long-term credit. These special credit institutions have as their primary aim the increase of the flow and the reduction of the cost of development finance, either to preferential areas or to priority sectors (for example, agriculture or research) or to small and medium-size business. In addition to this network of special credit institutions, there is a subsystem of credit under which the government shoulders part of the interest burden.

The bond market in Italy is well-developed. Mainly as a result of the special structure of government-sponsored institutions for development finance and subsidized interest rates, the growth of the capital market and stock exchanges was far less important than in other Western industrialized countries. The development of the stock exchange in Italy was initially hampered by the archaic structure and rules of the markets and by tax problems connected with the registration of shares. More recently, however, the market has been modernized; the Borsa Italiana, which manages the stock exchange, became operational in 1998.

TRADE

Italy has a great trading tradition. Jutting out deeply into the Mediterranean Sea, the country occupies a position of strategic importance, enhancing its trading potential not only with eastern Europe but also with North Africa and the Middle East. Italy has historically maintained active relations with eastern European countries, Libya, and the Palestinian peoples. These links have been preserved even at times of great political tension, such as during the Cold War and the First Persian Gulf War of 1991. Membership in the EC from 1957 increased Italy's potential for trade still further, giving rise to rapid economic growth. However, from that time, the economy was subject to an ever-widening trade deficit. Between 1985 and 1989 the only trading partner with which Italy did not run a deficit was the United States. Italy began showing a positive balance again in the mid-1990s. Trade with other EU members accounts for more than half of Italy's transactions. Other major trading partners include the United States, Russia, China, and members of the Organization of Petroleum Exporting Countries (OPEC).

Trade deficit

Italy's trading strength has traditionally been built on textiles, food products, and manufactured goods. During the second half of the 20th century, however, products from Italy's burgeoning metal and engineering sector, including automobiles, rose to account for a majority share of the total exports, which it still retains; they are followed by the textiles, clothing, and leather goods sector. The most avid customers of Italian exports are Germany, France, the United States, the United Kingdom, and Spain.

Italy's main imports are metal and engineering products, principally from Germany, France, the United States, and the United Kingdom. Chemicals, vehicle, and mineral imports are also important commodities. Italy is a major importer of energy, with much of its oil supply coming from North Africa and the Middle East.

Membership in the EC (now part of EU) was the most beneficial economic factor in Italian trade during the post-World War II period. The later accession of Greece, Spain, and Portugal to the group created stiff competition for Mediterranean agricultural products, especially fruit, wine, and cooking oil. At the beginning of the 21st century, however, the expanded EU and the weakness of the new euro currency allowed for export growth in Italy.

Euro currency

TRANSPORTATION

Water transport. Water transport was the first important means of linking Italy with its Mediterranean trading partners, even though its only navigable internal water is the Po River. At the time of unification in the 19th century, the ports of Venice, Palermo, and Naples were of great significance, and the Italian merchant fleet was preeminent in the Mediterranean Sea. The 4,600 miles (7,400 km) of Italian coastline are punctuated by many ports, and a large majority of imports and exports arrive and leave the country by sea. The principal dry-cargo ports are Venice, Cagliari, Civitavecchia, Gioia Tauro, and Piombino, while those handling chiefly petroleum products are Genoa, Augusta, Trieste, Bari, and Savona. Naples and Livorno handle both types of cargo. Half of the commercial port traffic is concentrated on only one-tenth of the coastline. The industries of Piedmont and Lombardy make heavy demands on the maritime outlets, particularly Genoa, which is the most extensive and important Italian port but which has great difficulty expanding because of the mountains surrounding it.

Rail transport. The main period of railway construction was about the time of unification, from 1860 until 1873. The heavy costs involved in laying down the infrastructure caused the government to sell off its stake in 1865. By this time the networks serving Milan, Genoa, and Turin in the north were well-developed. These were followed by links through the Po Valley to Venice; to Bari, along the Adriatic coast; down the Tyrrhenian coast, through Naples, to Reggio di Calabria; and from Rome to the Adriatic cities of Ancona and Pescara. The Sicilian and Sardinian networks also were built. A period of rationalization and modernization followed in 1905 when the network was renationalized; building of new rail lines continued throughout the 20th century. An exceptional feature was the early electrification of the lines, many of which ran through long tunnels and were ill-suited to steam power. This modernization was due to Italy's early development of hydroelectricity.

Although the rail network is well distributed throughout the peninsula, there are important qualitative differences between its northern and southern components. The north enjoys more frequent services, faster trains, and more double track lines than the south. Compared with other European networks, the Italian trains carry little freight but many passengers, partly because the railways failed to keep pace with the rapid rate of industrialization after World War II, while the passenger lines were made inexpensive

Differences between north and south

through government subsidies. Eighty percent of the rail network was controlled by the state via Ferrovie dello Stato ("State Railways") before it was privatized in 1992.

The Italian railways are connected with the rest of Europe by a series of mountain routes, linking Turin with Fréjus in France, Milan with Switzerland via the Simplon Tunnel, Verona to Austria and Germany via the Brenner Pass, and Venice to eastern Europe via Tarvisio. In the late 20th century, routes were expanded, extended, and modernized, including the addition of high-speed lines and computerized booking and freight control systems. The railway network extends some 12,000 miles (19,000 km).

Road transport. The Italian road network is subdivided into four administrative categories—express highways (*autostrade*) and national, provincial, and municipal roads (*strade statali*, *strade provinciali*, and *strade comunali*, respectively). Road construction in Italy flourished between 1955 and 1975. Between 1951 and 1980, surfaced roads, excluding highways and urban streets, increased by 72 percent to cover more than 183,000 miles. Automobile sales increased faster than in any other western European economy during this period. Much of this was due to mass production of cheap models by Fiat. Road construction in the south particularly benefited from funds released by the Southern Development Fund.

The highway system

More spectacular than general road construction was the development of the highway system. This project was farmed out to concessionary companies and financed by tolls, releasing it from the slow state bureaucracy and explaining its rapid progress. By the 1980s the network extended over 3,700 miles, making it second in Europe (only West Germany's was bigger). The main axis runs north-south from Chiasso on the Swiss border via Milan, Bologna, Florence, and Rome all the way south to Reggio di Calabria at the very tip of the peninsula. Another major route cuts southward from the Brenner Pass along the Adriatic coast to Bari and Taranto. A dense latticework of highways serves the north, linking Turin to Milan, Venice, and Trieste on an east-west axis and to Bologna and Genoa. Other east-west routes link Rome to Pescara across the Apennines and connect Naples to Bari. Commercial road transport has increased in recent years; Italy has one of the five largest trucking fleets in Europe.

Congestion is one of the main problems facing Italy's urban streets. Many town centres are based on medieval street plans and are unable to cope with levels of traffic and pollution generated by a population with one of the highest rates of automobile ownership in western Europe. Several cities, including Rome and Milan, have introduced measures to reduce the number of cars entering the city centres at peak hours and promoted other modes of transport; nonetheless, the proportion of the population commuting to work by car increased slightly in the 1990s.

Air transport. Of the small proportion of freight passing through Italian airports, a majority of it is processed either at Malpensa Airport near Milan or at Leonardo da Vinci Airport (in Fiumicino) near Rome. These airports, nearly equally, also handle the bulk of passenger traffic, though Linate Airport in Milan and Tessera Airport in Venice carry a large number as well. Many of the other regional airports (including those at Turin, Genoa, Verona, Bologna, Rimini, Pisa, Naples, Brindisi, Palermo, Catania, and Cagliari) are used for domestic flights, except during the peak tourist season, when they may absorb some of the vacation traffic from other European destinations.

The most frenetic developments in air transport occurred in the 1960s, with a tenfold increase in freight traffic and a sevenfold increase in passengers. At this time, Alitalia, Italy's national airline, became one of the largest in Europe. It remained viable by surviving the oil crisis of the 1970s, diversifying as a result of airline deregulation in the 1980s, and forming partnerships with foreign airlines in the 1990s and the early 21st century.

Telecommunications. In regard to its connectedness over Internet and telephone lines, Italy was on par with other western European nations in the 21st century. In addition, the growth of electronic commerce (e-commerce), particularly wireless services, benefited from the widespread use of cellular phones; Italy was one of the largest markets in Europe, with some nine-tenths of the population listed as mobile phone users. (R.L.K./P.E.S./Ed.)

Administration and social conditions

GOVERNMENT

Constitutional framework. The Italian state grew out of the kingdom of Sardinia-Piedmont, where, in 1848, King Charles Albert introduced a constitution that remained the basic law of his kingdom and, later, of Italy, for nearly 100 years. It provided for a bicameral Parliament with a Cabinet appointed by the king. With time, the power of the crown diminished, and ministers became responsible to Parliament rather than to the king. Although the constitution remained formally in force after the fascists seized power in 1922, it was devoid of all substantial value. After World War II, on June 2, 1946, the Italians voted in a referendum to replace the monarchy with a republic. A Constituent Assembly worked out a new constitution, which came into force on Jan. 1, 1948.

The constitution has built-in guarantees against easy amendment in order to make it virtually impossible to substitute for it a dictatorial regime. It is upheld and watched over by the Constitutional Court. The republican form of government cannot be changed. The constitution contains some preceptive principles, applicable from the moment it came into force, and some programmatic principles, which can be realized only by further enabling legislation.

The constitution is preceded by the statement of certain basic principles, including the definition of Italy as a democratic republic, in which sovereignty belongs to the people. Other principles concern the inviolable rights of man, the equality of all citizens before the law, and the obligation of the state to abolish social and economic obstacles that limit the freedom and equality of citizens and hinder the full development of individuals.

Personal freedom

The constitution guarantees many forms of personal freedom: the privacy of correspondence; the right to travel at home and abroad; the right of association for all purposes that are legal, except in secret or paramilitary societies; and the right to hold public meetings, if these are consistent with security and public safety. There is no press censorship, and freedom of speech and writing is limited only by standards of public morality. The constitution stresses the equality of the spouses in marriage and the equality of their children to each other. Family law has seen many reforms, including the abolition of the husband's status as head of the household and the legalization of divorce and abortion.

One special article in the constitution concerns the protection of linguistic minorities. Regarding religion, the constitution establishes the liberty of all religions before the law but also recognizes the special status granted the Roman Catholic church by the Concordat of 1929. That special status was modified and reduced in importance by a new agreement between church and state signed in 1984. The church renounced much of its tax-exempt status and much of its jurisdiction over the annulment of church marriages. Religious weddings must conform to civil legal standards. Religious instruction in public schools is offered only on request. Salaries to the Catholic clergy and other forms of financial support for the church have been phased out. Owing to these changes and to the liberal tendencies manifested by the church after the Second Vatican Council in the 1960s, religion is much less a cause of political and social friction in contemporary Italy than it was in the past.

The Constitutional Court

The constitution is upheld by the Constitutional Court, which is composed of 15 judges, of whom 5 are nominated by the president of the republic, 5 by Parliament, and 5 by judges from their own ranks. Members must have certain legal qualifications and experience. The term of office is nine years, and Constitutional Court judges are not eligible for immediate reappointment. The court performs four major functions. First, it judges the constitutionality of state and regional laws and of acts having the force of law. Second, the court resolves jurisdictional conflicts between ministries or administrative offices of the central government or between the state and a particular region or between two regions. Third, it judges indictments instituted by Parliament. When acting as a court of indictment, the 15 Constitutional Court judges are joined by 16 additional lay judges chosen by Parliament. Fourth, the court determines whether or not it is permissible to hold referenda on particular topics. The constitution specifically excludes from the field of referenda financial decisions, the granting of amnesties and pardons, and the ratification of treaties.

The legislature. Parliament is bicameral and comprises the Chamber of Deputies and the Senate, both elected by popular vote and with equal powers. In theory, the Senate should represent the regions and in this way differ from the lower chamber, but in practice the only real difference between them lies in the minimum age required for the electorate and the candidates: 18 and 25, respectively, for deputies and 25 and 40 for senators. Deputies and senators alike are elected for a term of five years, which can be extended only in case of war. Parliamentarians cannot be penalized for opinions expressed or votes cast, and deputies or senators are not obligated to vote according to the wishes of their constituents. Unless removed by parliamentary action, deputies and senators enjoy immunity from arrest, criminal trial, and search. Their salary is established by law, and they qualify for a pension.

Both houses are officially organized into parliamentary parties. Each house also is organized into standing committees, which reflect the proportions of the parliamentary groups. However, the chairmanship of parliamentary committees is not the exclusive monopoly of the majority. Besides studying bills, these committees act as legislative bodies. The new parliamentary rules have followed the United States' pattern and have given the standing committees extensive powers of control over the government and administration. All these features explain why the government has a limited ability to control the legislative agenda and why parliamentarians are often able to vote contrary to party instructions and to avoid electoral accountability. The abolition of secret voting on most parliamentary matters at the end of the 1980s did not significantly change this situation.

Special majorities are required for constitutional legislation and for the election of the president of the republic, Constitutional Court judges, and members of the Superior Council of the Magistrature. The two houses meet jointly to elect and swear in the president of the republic and to elect one-third of the members of the Superior Council of the Magistrature and one-third of the judges of the Constitutional Court. They may also convene to impeach the president of the republic, the president of the Council of Ministers, or individual ministers.

Each year, the annual budget and the account of expenditure for the past financial year are presented to Parliament for approval. The budget, however, does not cover all public expenditure, nor does it include details of the budgets of many public bodies, over which, therefore, Parliament has no adequate control. International treaties are ratified by means of special laws.

The process of legislation

The most important function of Parliament is ordinary legislation. Bills may be presented in Parliament by the government, by individual members, or by bodies such as the National Council for Economy and Labour, various regional councils, or communes, as well as by petition of 50,000 citizens of the electorate or through a referendum. Bills are passed either by the standing committees or by Parliament as a whole. In either case, the basic procedure is the same. First, there is a general debate followed by a vote; then, each of the bill's separate articles is discussed and voted on; finally, a last vote is taken on the entire bill. All bills must be approved by both houses before they become law; thus, whenever one house introduces an amendment to the draft approved by the other house, the latter must approve the amended draft. The law is then promulgated by the president of the republic unless he considers it unconstitutional or inappropriate; in that case, he remands the bill to Parliament for reconsideration. If the bill is, nevertheless, passed a second time, the president is obliged to promulgate it. The law comes into force when published in the *Gazzetta Ufficiale*.

The presidential office. The president of the republic is irremovable, and his seven years of office cannot be shortened. He may, however, be impeached for high treason or offenses against the constitution, even while in office. He is elected by a college comprising both chambers of Parliament, together with three representatives from every region. The two-thirds majority required guarantees that the president is acceptable to a sufficient proportion of the populace and the political partners. The minimum age for presidential candidates is 50. If the president is temporarily unable to carry out his functions, the president of the Senate acts as his deputy. If the impediment is permanent or if it is a case of death or resignation, a presidential election must be held within 15 days.

Special powers and responsibilities are vested in the president of the republic. He promulgates laws and decrees having the force of law, calls special sessions of Parliament, delays legislation, authorizes the presentation of government bills in Parliament, and, with parliamentary authorization, ratifies treaties and declares war. However, some of these acts are duties that must be performed by him, whereas others have no validity unless countersigned by the government. The president commands the armed forces and presides over the Supreme Council of Defense and the Superior Council of the Magistrature. He may dissolve Parliament either on his own initiative (except during the last six months of his term of office), having consulted the presidents of both chambers, or at the request of the government. He may appoint 5 lifetime members of the Senate and appoints 5 of the 15 Constitutional Court judges. He also appoints the president of the Council of Ministers, the equivalent of a prime minister. Whenever a government is defeated or resigns, it is his duty, after consulting eminent politicians and party leaders, to appoint the person most likely to win the confidence of Parliament; this person is usually designated by the majority parties, and the president has limited choice. However, in more recent times, the powers of the president have de facto increased.

The government. The government comprises the president of the Council of Ministers and the various other ministers responsible for particular departments. Ministerial appointments are negotiated by the parties constituting the government majority. Each new government must receive a vote of confidence in both houses of Parliament within 10 days of its appointment. If at any time the government fails to maintain the confidence of either house, it must resign. Splits in the coalition of two or more parties that had united to form a government have caused most of the resignations of governments. According to the constitution, the president of the Council of Ministers is solely responsible for directing government policies and coordinating administrative policy and activity. In reality, the president tends to function as a negotiator between government parties and factions. The government can issue emergency decree laws signed by the president of the republic, provided such laws are presented to Parliament for authorization the day they are issued and receive its approval within 60 days. Without such approval they automatically lapse. The government and, in certain cases, individual ministers issue administrative regulations and provisions, which are then promulgated by presidential decree.

President of the Council of Ministers

Regional and local government. The republic is divided into regions, provinces, and communes. There are 15 ordinary regions and an additional 5 (Sicily, Sardinia, Trentino–Alto Adige, Friuli–Venezia Giulia, and Valle d'Aosta) to which special autonomy has been granted. The regions with ordinary powers are Piedmont, Lombardy, Veneto, Liguria, Emilia-Romagna, Tuscany, Umbria, Marche, Latium, Abruzzi, Molise, Campania, Puglia, Basilicata, and Calabria. Italy can thus be considered a regional state. The modern regions correspond to the traditional territorial divisions. The powers of the five special regions derive from special statutes adopted through constitutional laws. The organs of regional government are the regional council, a popularly elected deliberative body with power to pass laws and issue administrative regulations; the regional committee, an executive body elected by the council from among its own members; and the president of the regional committee. The regional committee and its president are required to resign if they fail to retain the confidence of the council. Voting in the regional councils is rarely by secret ballot.

Participation in national government is a principal function of the regions: regional councils may initiate parliamentary legislation, propose referenda, and appoint three delegates to assist in presidential elections. With regard to regional legislation, the five special regions have exclusive competence in certain fields—such as agriculture, forestry, and town planning—while the ordinary regions have competence over them within the limits of fundamental principles established by state laws. The legislative powers of both special and ordinary regions are subject to certain constitutional limitations, the most important of which is that regional acts may not conflict with national interests. The regions can also enact legislation necessary for the enforcement of state laws when the latter contain the necessary provisions. The regions have administrative competence in all fields in which they have legislative competence. Additional administrative functions can be delegated by state laws. The regions have the right to acquire property and the right to collect certain revenues and taxes.

The state has powers of control over the regions. The validity of regional laws that are claimed to be illegal can be tested in the Constitutional Court, while those considered inexpedient can be challenged in Parliament. State supervisory committees presided over by government-appointed commissioners exercise control over administrative acts. The government has power to dissolve regional councils that have acted contrary to the constitution or have violated the law. In such an event, elections must be held within three months.

Communes

The organs of the commune, the smallest local government unit, are the popularly elected communal council, the communal committee, or executive body, and the mayor. The communes have the power to levy and collect limited local taxes, and they have their own police, although their powers are much inferior to those exercised by the national police. The communes issue ordinances and run certain public health services, and they are responsible for such services as public transportation, garbage collection, and street lighting. Regions have some control over the activity of the communes. Communal councils may be dissolved for reasons of public order or for continued neglect of their duties.

The organization of the provinces, units midway in size between regions and communes, is analogous to that of the communes; they each have councils, committees, and presidents.

There are certain central government officials whose duties lie in the sphere of local government. These include the government commissioner of each region, who supervises the administrative functions performed by the state and coordinates them with those performed by the region; the prefect, resident in each province, who is responsible for enforcing the orders of the central government and has powers of control over the organs of the province and communes; and the *questore*, who is the provincial chief of the state-run police. Certain local government officials also have central government duties: among them are the president of the regional committee who, in directing the administrative functions that the state delegates to the region, performs a specific state duty; and the mayor of a commune who, in his capacity as an agent of the central government, registers births, deaths, marriages, and migrations, maintains public order (though in practice this is dealt with by the national police), and can, in cases of emergency, issue ordinances concerning public health, town planning, and the local police.

The political process. *Elections.* For almost half a century after World War II, Italy's electoral system was based on proportional representation, a system in which seats in an elected body are awarded to political parties according to the proportion of the total vote that they receive. Between 1993 and 1995, however, several changes were made by national legislation and popular referenda. On the national level, the Chamber of Deputies and the Senate are now elected by a combination of proportionality and plurality. Seventy-five percent of the seats in these two chambers are now filled from single-member districts by individual candidates who win the largest number of votes in each district. The other 25 percent of the seats are awarded to candidates from party lists on a proportional basis. The number of votes obtained by the winner in single-member districts is fully (for senators) or partially (for deputies) subtracted before allocating proportional seats, thus introducing a further element of proportionality.

In regional elections, voters cast two ballots. The first is cast in a contest for 80 percent of the seats in the regional council, which are awarded on a proportional basis. The second ballot is employed in a plurality vote; the regional coalition that wins a plurality is awarded all the remaining seats as well as the presidency of the regional government. Split voting is allowed.

In provincial elections, only one vote is cast. If a single provincial list wins more than 50 percent of the votes, seats are divided among all the lists according to their proportion of the vote, and the presidency goes to the head of the winning list. Otherwise, a run-off election must take place between the two most successful lists, with the winner taking 60 percent of the seats.

A similar system is employed in municipal elections in cities with more than 15,000 inhabitants. In this case, however, two ballots are cast, one for mayor and one for the council. Split voting is permitted. In smaller cities only one ballot is cast; the winning list is awarded two-thirds of the seats as well as the mayoralty.

Political parties. From the end of World War II until the 1990s, Italy had a multiparty system with two dominant parties, the Christian Democratic Party (Partito della Democrazia Cristiana; DC) and the Italian Communist Party (Partito Comunista Italiano; PCI), and a number of small yet influential parties. The smaller parties ranged from the neofascist Italian Social Movement (Movimento Sociale Italiano; MSI) on the right to the Italian Socialist Party (Partito Socialista Italiano; PSI) on the left; a number of small secular parties occupied the centre. The DC, in various alliances with smaller parties of the centre and left, was the dominant governing party, and the principal opposition parties were the PCI and the MSI.

The postwar party system described above was radically altered by the fall of communism, by a wave of judicial prosecutions of corrupt officials that involved most Italian political parties, and finally by the electoral reforms of the 1990s. The DC, riven by scandal, was replaced by a much smaller organization, the Italian Popular Party (Partito Popolare Italiano; PPI), which virtually disappeared after elections in 1994. By that time three new parties had arisen to dominate the political right: Forza Italia (FI; loosely translatable as "Go Italy"), a vaguely neoliberal alliance created in 1994 by the media tycoon Silvio Berlusconi; the Northern League (Lega Nord; LN), formed in 1991, a federalist and fiscal-reform movement with large support in the northern regions; and the National Alliance (Alleanza Nazionale; AN), which succeeded the MSI in 1994 but whose political platform renounced its fascist past. Meanwhile, the PCI remained an important electoral force under a new name, the Democratic Party of the Left (Partito Democratico della Sinistra; PDS). Thus, the Italian political spectrum, which had previously been dominated by parties of the centre, became polarized between parties of the right and left.

Trade unions. The constitution establishes the right to organize trade unions. The right to strike is guaranteed by the constitution and remains a very potent weapon in the hands of the trade unions. Three major labour federations exist, each close to different political areas: the Italian General Confederation of Labour (Confederazione Generale Italiana del Lavoro; CGIL; closest to the left), the Italian Confederation of Free Workers (Confederazione Italiana di Sindicati Liberi; CISL; closest to the Catholic movement), and the Italian Labour Union (Unione Italiana del Lavoro; UIL; closest to the secular parties). A number of independent unions are also active, especially in the public-service sector. They increasingly challenge the monopoly of the three confederations on national contractual negotiations and are quite militant.

The participation of the citizen. All citizens 18 years and over may vote. The turnout for elections in Italy is high, reaching well over 80 percent of the electorate for parliamentary elections. Citizens may also subscribe to national referenda or petitions designed to abrogate a law or an executive order; such a petition must be signed by 500,000 members of the electorate or sponsored by five regional councils. Abrogative referenda have been used extensively since the 1970s to make possible a wide range of institutional and civic reforms. Abrogative referenda are provided for with regard to all regional legislation, and some regions have a provision for holding ordinary referenda. The constitution also provides that 50,000 members of the electorate may jointly present to Parliament a draft bill.

Voting

JUSTICE

The Italian judicial system consists of a series of courts and a body of judges who are civil servants. Judges and prosecutors belong to the same civil service career, and their positions are interchangeable. The judicial system is unified, with every court being part of the national network. The highest court in the central hierarchy is the Supreme Court of Cassation; it has appellate jurisdiction and gives judgments only on points of law. The 1948 constitution prohibits special courts with the exception of administrative courts and courts-martial, although a vast network of tax courts has survived from an earlier period. The administrative courts have two functions: the protection of *interessi legittimi*—that is, the protection of individual interests directly connected with public interests and protected only for that reason—and the supervision and control of public funds. Administrative courts are also provided by the judicial sections of the Council of State, the oldest juridical-administrative advisory organ of government. The Court of Accounts has both an administrative and a judicial function; the latter involves primarily fiscal affairs. The Superior Council of the Magistrature, provided for by the constitution and intended to guarantee the independence and integrity of the judiciary, was formed only in 1958. It attends to the careers, assignments, and disciplining of judges. Two-thirds of its members are elected by the judges and one-third by Parliament. The president and the Public Prosecutor of the Court of Cassation also belong to it. Elections tend to politicize the council, which has become an influential force in Italian politics.

Italian law is codified and based on Roman law, in particular as regards civil law. The codes of the Kingdom of Sardinia in civil and penal affairs, derived from the French Napoleonic model, were extended to the whole of Italy when unification was achieved in the mid-19th century. In the period between World Wars I and II, these codes were revised. The Constitutional Court has declared a number of articles unconstitutional. The revised 1990 penal code replaced the old inquisitory system with an accusatory system akin to that of common-law countries. Besides the codes, there are innumerable statute laws that integrate the codes and regulate areas of law, such as public law, for which no codes exist.

The constitution stresses the principle that the judiciary should be independent of the legislature and the executive. For this reason jurisdictional functions can be performed only by ordinary magistrates, and extraordinary tribunals may not be set up. Judges cannot be dismissed, they are not subject to hierarchical superiors, and their careers rest on seniority.

ARMED FORCES

The armed forces are commanded by the president of the republic, who also presides over the Supreme Council of Defense, comprising the president of the Council of Ministers; the ministers of defense, the interior, foreign affairs, industry, and the treasury; and the chief of defense general staff. Military service for men is obligatory. Although the constitution specifies that the armed forces must embody the democratic spirit of the republic, their activity is free from any political control. Italy's membership in the North Atlantic Treaty Organization (NATO) since 1949 has given the allied command a certain degree of control over the Italian forces.

There are two police forces in Italy with general duties: the Pubblica Sicurezza, which is under the authority of the home secretary, and the Carabinieri, a corps of the armed forces, under the minister of defense. The functions of the police are the prevention, suppression, and investigation of crimes. All functions are performed by both police forces. In the year following World War II, the Constitutional Court and Parliament removed the special prerogatives that fascism had granted the police in their preventive and suppressive duties. When engaged in the investigation of crimes, the police are placed by the constitution under the authority of the courts; however, the actual subordination of the two forces to two different government ministries is a source of conflict with regard to their technical subordination to the judiciary. In addition to these two police forces, there are special police for customs and for excise and revenue. There are also communal police and prison guards as well as private police that operate in a limited field under the supervision of the regular police.

Two police forces

EDUCATION

The constitution guarantees the freedom of art, science, and teaching. It also provides for state schools and guarantees the independence of the universities. Private schools (mainly run by religious bodies) are permitted. The constitution further states that the public schools are open to all and makes provision for scholarships and grants. Education is compulsory only for those aged 6 to 14 years. The school system begins with kindergarten for the 3- to 6-year-olds. Elementary schools are attended by children between the ages of 6 and 11, at which stage most go on to secondary schools for 11- to 14-year-olds, but those wishing to study music go directly to the conservatories. Postsecondary schooling is not compulsory and includes a wide range of technical and trade schools, art schools, teacher-training schools, and scientific and humanistic preparatory schools. Pupils from these schools can then continue their education at universities, where graduate studies beyond the masters' level have been introduced only in recent years.

Approximately 2 Italians in 100 are illiterate. Of the total population above the age of six, close to 4 percent have a college degree and 18 percent a high-school diploma, and close to 31 percent attended junior high school. But educational attainment is higher in the younger generations. Twenty-nine percent of people of university age attend university, and 73 percent of people of high-school age attend high school. Most schools and universities are run by the state, with programs that are uniform across the nation. About 14 percent of students attend private schools. University fees are low, and enrollment is unrestricted for most students with a postsecondary school diploma.

HEALTH AND WELFARE

Social services. Italy possesses an extensive social security and welfare system that provides coverage for the great majority of the population. The system is run by a sprawling number of state agencies that supervise all social services, make available benefits in the case of accident, illness, disability, or unemployment, and provide assistance for the elderly. The largest of these agencies, which administers a wide range of benefits, is the National Social Insurance Institute (Istituto Nazionale della Previdenza Sociale; INPS).

A comprehensive national health service and national medical insurance based on Local Medical Units (Unità Sanitarie Locali; USL) were created in 1978. The funds of these agencies, coming largely from the state, are considerable, but, owing to the politicization of the agencies and to increasing costs, the services they provide are often inadequate. (G.DiP.)

Cultural life

The 20th century saw the transformation of Italy from a highly traditional, agricultural society to a progressive, industrialized state. Although the country was politically unified in 1861, regional identity remains strong, and the nation has developed unevenly as a cultural entity.

DAILY LIFE

The changing role of women

Since World War II, Italian society has profoundly changed, with a significant impact on daily life. One of the main elements of change is the more visible role women play in society outside the home, such as increased participation in higher education and the professions. One aspect of this change is that Italy has one of the lowest average numbers of children per woman in the world, as well as one of the lowest birth rates. Nevertheless, women engage in more domestic work than men, and they have continuing responsibilities in the care of the family.

For Italian families, among the most popular daily leisure activities are watching television, listening to the radio, reading newspapers, and going to the cinema; reading books and engaging in sports are less common for most people. According to surveys, Italians are very satisfied with their family relations, friendships, and health status, while their economic status and their working positions are less satisfactory. This is especially the case in southern Italy, where there are fewer job opportunities and where unemployment is high.

Though the popularity of home entertainment has grown, the use of public spaces remains important. Young Italians meet friends on a daily basis, often in the cities' piazzas in the evenings, and make frequent trips to bars, cinemas, pizzerias, and discos. Coastal areas are popular summer destinations. The automobile retains a strong hold on daily life as well. Ownership levels are high, and many cities and towns suffer congestion and pollution as a result.

Food

Food is traditionally a primary element of Italian life. Work patterns in the country revolve around the midday meal, and the Italian bars cater cheaply and quickly to the casual diner. Although there is in Italian cuisine a wide variety of regional differences, Italians share a love of pasta, which bears such euphonious names as spaghetti ("little strings"), penne ("feathers"), macaroni ("little dear things"), and orecchiette ("little ears"). In the south, pasta is often dressed with sauces made of olive oil, tomatoes, and spices; in the north, especially in Piedmont, it is coated in cream, butter, and cheese. Many foreigners have grown accustomed to these regional variations, as Italian cuisine has become a popular cultural export.

For most Italians in the 21st century, religious activity plays a much smaller role in daily life than it did in the prior century and is usually concentrated on Sundays or on special celebrations such as Christmas and Easter. However, older generations, especially in rural settlements, tend to be more involved and may attend mass every day.

Regional life in Italy is typified by a diversity of customs and a variety of festivals, even if it is their appeal to the tourist industry that maintains them. The majority of religious festivals are Roman Catholic, dedicated to the Madonna or to different saints. The feast of the Epiphany on January 6 exemplifies religious diversification. Traditionally, a witch called the Befana brings gifts to children on this day. However, in the villages of Mezzojuso and Piana degli Albanesi, both near Palermo, the Epiphany is celebrated according to the Byzantine and Albanian rites, respectively. The most notable Carnival celebrations are held at Viareggio and Venice, where in 1992 they were financed for the first time by major sponsors.

Italy's strong agricultural tradition gives rise to a multitude of festivals celebrating the harvest, food, country, and seafaring pursuits. These festivals reflect the traditional activities of the area in which they are held. For example, the olive and *bruschetta* festival at Spello (near Perugia) marks the end of the olive harvest, the fish festival at Termoli reflects the fishing tradition of the port, just as the hazelnut festival in Canelli (near Asti) gives testimony to the importance of this local crop. At Senale (near Bolzano) the traditional migration of sheep across the Giorgio glaciers is celebrated, while fishermen in the port of Aci Trezza (near Catania) stage a farcical swordfish hunt every June.

Some festivals are more sporting in nature, such as the historic horse race the Corsa del Palio in Siena, Florence's "football match" in 16th-century costume, and the regattas of Venice, while others commemorate historical events, such as the Lily Festival at Nola (near Naples), recalling the return of St. Paulinus of Nola in 394 after a long imprisonment in Africa, and the festival of Piedigrotta in Naples, commemorating the battle of Velletri in 1744.

THE ARTS

Primacy in the arts

Italy was at the forefront of the artistic and intellectual developments of the Renaissance, which drew their impetus from a reappraisal of the Classical Greek and Roman world. Artists and scholars in Italy were especially well placed to take the lead in such a revival, because they were surrounded by the material remains of antiquity. Earlier Romanesque and Gothic forms in both art and architecture were supplanted by the Renaissance, which escalated with a flourish into the Baroque styles of the 16th century.

Visual arts. The great names in Italian art through the centuries make a long list that includes, among many others, Giotto, Donatello, Brunelleschi, Michelangelo, Leonardo da Vinci, Titian, Bernini, and Tiepolo. Broadly characterized by a warmth of colour and light, Italian painting enjoyed preeminence in Europe for hundreds of years. Continuous subjection to foreign powers, however, eventually enfeebled Italy's artistic contribution, which sank into provincialism. Ties with European art were renewed about 1910 by the work of the Futurists, led by the poet Filippo Marinetti and the painters Umberto Boccioni and Giacomo Balla. Futurism was succeeded by the Metaphysical paintings of Giorgio de Chirico, who influenced the Surrealists until the 1920s, when he began to produce more traditional canvases. Since the death of Giorgio Morandi in 1964, his subtle, quietist paintings have placed him in increasingly high regard. Argentinian-born Lucio Fontana's work exemplified the modern artist's quest for form, expressed, for example, by a blank canvas slashed open by a knife. Modern additions to the Italian tradition of sculpture included the works of Giacomo Manzù, Gio Pomodoro, Marino Marini, Luciano Minguzzi, Alberto Viani, Harry Bertoia, Mirko Basaldella, and Emilio Greco.

Architecture. The traditional image of old Italian towns situated around piazzas adorned with fountains remains valid in a country where ruins from Classical antiquity may stand alongside modern construction marvels. The Rationalist architecture movement of 1926 produced one of the outstanding Italian architect-engineers of the 20th century, Pier Luigi Nervi, architect of the Turin exhibition complex and the UNESCO headquarters in Paris. Marcello Piacentini was responsible for much of the imposing architecture of the fascist period, such as the Esposizione Universale di Roma (EUR) area in Rome. Innovative architecture is represented in Milan's Marchiondi Spagliardi Institute, by Vittoriano Viganò. Other architects of note include Renzo Piano, known for his international museums; Aldo Rossi, whose critical writings rivaled his built works; and Paolo Portoghesi, who created public buildings from curvilinear forms.

Literature. Italian literature, and indeed standard Italian, have their origins in the 14th-century Tuscan dialect—the language of its three founding fathers, Dante, Petrarch, and Boccaccio. The thread of literature bound these pioneers together with later practitioners, such as the scientist and philosopher Galileo, dramatist Carlo Goldoni, lyric poet Giacomo Leopardi, Romantic novelist Alessandro Manzoni, and poet Giosuè Carducci.

After the unification of Italy, writers began to explore subjects theretofore considered too lowly for literary consideration, such as poverty and living conditions in the Mezzogiorno. Writers such as Giovanni Verga invented a new vocabulary to give expression to them. Among the first generation of women writers was a Sardinian, Grazia Deledda, who won the 1926 Nobel Prize for Literature. However, the most prominent Italian woman writer of the 20th century was Elsa Morante.

The themes of writers in the 20th century ranged widely.

The flamboyant patriotism of Gabriele d'Annunzio in the early decades of the century gave way to the existentialist concerns of Deledda and Ugo Ojetti, who focused on local aspects of Italian life. The fascist period forced many writers underground but at the same time provided inspiration for their work, as in the case of Ignazio Silone and Carlo Levi. Italo Svevo and Luigi Pirandello pioneered the psychoanalytic literary genre, prior to the revival of realism by writers such as Elio Vittorini. Alberto Moravia wrote of the corruption of the upper middle classes and gained notoriety for the eroticism of his narrative.

By the 1960s the literary world joined the protest movement against the corruption of the state, and poetry eclipsed the novel as the primary literary genre. Pier Paolo Pasolini, a poet, critic, and filmmaker, was the dominant creative figure of the period. Eugenio Montale and Salvatore Quasimodo won Nobel Prizes for their poetry, and Giuseppe Ungaretti founded Hermeticism.

Of literature in the late 20th century, the work of Italo Calvino, Umberto Eco, and Primo Levi met with much success abroad; within Italy the work of Cesare Pavese, Carlo Emilio Gadda, and Leonardo Sciascia was also well received. The last decades of the century saw the revival of the narrative and the historical novel, together with new forms of experimental and innovative language. In 1997 Dario Fo, a playwright known for his improvisational style, won the Nobel Prize for Literature.

Music. Italian music has been one of the supreme expressions of that art in Europe: the Gregorian chant, the innovation of modern musical notation in the 11th century, the troubadour song, the madrigal, and the work of Palestrina and Monteverdi all form part of Italy's proud musical heritage, as do such composers as Vivaldi, Alessandro and Domenico Scarlatti, Rossini, Donizetti, Verdi, Puccini, and Bellini.

Music in contemporary Italy, though less illustrious than in the past, continues to be important. Italy hosts many music festivals of all types—classical, jazz, and pop—throughout the year. In particular, Italian pop music is represented annually at the Festival of San Remo. The annual Festival of Two Worlds in Spoleto has achieved world fame. The state broadcasting company, Radiotelevisione Italiana (RAI), has four orchestras, and others are attached to opera houses; one of the best is at La Scala in Milan. The violinists Uto Ughi and Salvatore Accardo and the pianist Maurizio Pollini have gained international acclaim, as have the composers Luciano Berio, Luigi Dallapiccola, and Luigi Nono.

La Scala opera house

Contemporary productions maintain Italy's eminence in opera, notably at La Scala in Milan, as well as at other opera houses such as the San Carlo in Naples and La Fenice Theatre in Venice, and the annual summer opera productions in the Roman arena in Verona. Tenors Luciano Pavarotti and Andrea Bocelli were among Italy's most acclaimed performers at the turn of the 21st century.

Theatre. There are a large number of theatres in Italy, many of which are privately run. The 15 publicly operated permanent theatres are funded by the state and supervised by the Ministry for Tourism. Three public organizations to promote theatrical activity in Italy are the Italian Theatre Board (Ente Teatrale Italiano; ETI), the Institute for Italian Drama (Istituto Dramma Italiano; IDI), concerned with promoting Italian repertory, and the National Institute for Ancient Drama (Istituto Nazionale del Dramma Antico; INDA). In 1990 the government tightened its legislation on eligibility for funding, which severely affected fringe and experimental theatres. Financial constraints in recent years have led to an increasing number of international coproductions.

Italian theatre has been active in producing outstanding contemporary European work and in staging important revivals. Although no native playwright has produced works that can rival those of Luigi Pirandello from the early 20th century, in the late 20th century Dario Fo received international acclaim for his highly improvisational style.

Film. The heyday of the Italian film was in the 1950s. Neorealism, best represented in the work of Roberto Rossellini and Vittorio De Sica, diverged from the escapism favoured during the interwar years to take a candid look at prevailing conditions in postwar Italy. This new style attracted world attention. Cinecittà, the complex of film studios built by Mussolini near Rome, became known as the Hollywood of Europe. Rome became the centre for the international jet set, who frequented the grand hotels and smart cafés of the Via Veneto, attracting a new breed of celebrity-hungry photographers known as paparazzi.

Neorealism

Federico Fellini propagated this image of the capital in films such as *Roma* (1972) and *La dolce vita* (1960; "The Sweet Life"). Pier Paolo Pasolini, on the other hand, took a grittier look at the Italian underworld in films such as *Accattone* (1961; *The Beggar*). Other directors who made a lasting contribution to the cinema of the day were Luchino Visconti, with masterpieces such as *Morte a Venezia* (1971; *Death in Venice*); brothers Paolo and Vittorio Taviani; and the screenwriter Cesare Zavattini. Some directors, such as Michelangelo Antonioni, Franco Zeffirelli, and Fellini, enjoyed more success abroad than at home.

More recently, Italian cinema fell into recession. Nevertheless, Italy can still claim some major international successes, including Bernardo Bertolucci's *The Last Emperor* (1987), Giuseppe Tornatore's *Cinema Paradiso* (1990), Gabriele Salvatores's *Mediterraneo* (1991), and Michael Radford's *Il postino* (1994; *The Postman*). Other directors of note are Gianni Amelio and Roberto Benigni, who won the Academy Award for best actor for a film he directed, *La vita è bella* (1997; *Life Is Beautiful*), which also won for best foreign movie. Italian films are increasingly coproductions of cinema and television companies. The Radiotelevisione Italiana (RAI) and Fininvest are presently Italy's largest film producers, accounting for more than half of the film output.

CULTURAL INSTITUTIONS

Italy's cultural heritage is an inescapable presence. The south and centre of Italy abound in vestiges of Greek and Etruscan civilization, and substantial Roman remains are visible throughout the peninsula. The most notable examples are the ancient Roman towns of Pompeii and Herculaneum near Naples and the remains in Rome itself. A wealth of monuments, churches, and palaces testify to Italy's cultural past, and the contents of its museums and galleries number more than 35 million pieces.

A statistical analysis of Italy's cultural institutions carried out by the presidency of the Council of Ministers shows that, as far as museums, libraries, and theatres are concerned, the cultural wealth of the nation is unevenly divided between the northern, central, and southern regions. More than half of Italy's art treasures are located in the centre, and almost all of the remainder is in the north.

Italy contains many historic places designated by the United Nations Educational, Scientific and Cultural Organization (UNESCO) as World Heritage sites. Among the places officially noted are the old city centres in Ferrara, Pienza, San Gimignano, Siena, and Urbino; archaeological sites in Agrigento, Aquileia, and Valcamonica; and the whole of the Amalfi coast and the Eolie Islands.

Museums and galleries. Italy's museums contain some of the most important collections of artifacts from ancient civilizations. The permanent collection in the National Museum in Taranto provides one of the most important insights into the history of Magna Graecia, while the archaeological collections in the Roman National Museum in Rome and in the National Archaeological Museum in Naples are considered among the best in the world. The same may be said of the Etruscan collection in the National Archaeological Museum of Umbria in Perugia, the Classical sculptures in the Capitoline Museums in Rome, and the Egyptian collection in the Egyptian Museum in Turin.

Italy's towering artistic achievement during the Renaissance is reflected in the magnificent collections in the Uffizi Gallery, the National Museum of the Bargello, and the Pitti Palace galleries in Florence. In addition to the Old Masters, the Uffizi, a public gallery since 1765, contains masterpieces by Michelangelo, Leonardo da Vinci, Botticelli, Piero della Francesca, Giovanni Bellini, and Titian. The Bargello holds a superb collection of Florentine sculpture, with works by Michelangelo, Cellini, Donatello, and

"The Last Judgment" by Michelangelo, 1534–41; in the Sistine Chapel, Vatican City.
Alberto Pizzoli/Corbis Sygma

the Della Robbia family. The Pitti Palace houses an impressive collection of paintings by Raphael, together with about 500 important works of the 16th and 17th centuries collected by the Medici and Lorraine families.

Many of Italy's major galleries are concerned primarily with their own regional heritage. For example, the Brera Art Gallery in Milan is rich in work from the northern Italian Lombard school; the Gallery of the Academy of Fine Arts in Venice is the major exponent of Venetian painting, as is the National Art Gallery in Siena of the Sienese school. The Vatican Palace and Vatican Museums, in Vatican City, are noted above all for the frescoes by Michelangelo in the Sistine Chapel, which were restored in the 1980s in one of the most ambitious conservation projects in Europe.

Sistine Chapel

A quarter of Italy's museums belong to the Italian state, just under half to local authorities, and a small proportion to public bodies, religious organizations, and private owners. The numbers of museum visitors are dependent on tourist figures. In 2000 there was an increase of tourists linked to the Jubilee celebrations of the Catholic church.

Libraries. Italy's national library system is controlled by the Central Office for Books, Manuscripts, and Cultural Institutes. This body oversees the work of cataloging and conserving the nation's books and directly controls the State Record Library and some 50 state libraries. The two principal national libraries are based in Rome and Florence. Their work is supported by the main national libraries of Bari, Naples, and Milan and their provincial branches.

Cultural institutes. Academies and societies, representing a multitude of interests, have proliferated in Italy. Indeed, academies of the fine arts had their origins in Italy; for example, the Academy of Fine Arts of Florence was founded as the Academy of Arts of Design in 1563, and the academy of Perugia dates to 1573. Rome's Academy of San Luca was a guild of painters, founded in 1577. Italy's most famous learned society is the National Academy of Lincei, of which Galileo was once a member. The most distinguished literary society is the Academy of Crusca, founded in Florence in 1582. There are also many historical and scientific societies, including the Cimento Academy, which opened in Florence in 1657. Foreign schools, which were established for the study of Italian art and culture, contribute significantly to academic life.

RECREATION

For a country in which only a small percentage of the population is actively involved in sports, Italy has produced an impressive number of champions in cycling, skiing, basketball, water polo, volleyball, and football (soccer). Especially popular is football, which some Italian scholars claim was invented in 16th-century Italy as *calcio* and introduced at the Palio festivals of Florence and Siena. Italian football teams excelled in international play in the 1930s and from the late 1960s onward. The national team has won the World Cup three times, most recently in 1982.

Automobile racing also is popular in Italy, and Italian engineers and drivers have contributed much to the sport. The Ferrari series of racing cars, first manufactured in 1946, have won more than 5,000 major races and set world records, as has the rival high-performance car Maserati.

Italian athletes have participated in every modern Olympiad. The Alpine town Cortina d'Ampezzo hosted the 1956 Winter Olympics, and the 1960 Summer Games were held in Rome. Perhaps Italy's most famous Olympian, diver Klaus Dibiasi won three consecutive gold medals in platform diving in the 1960s and '70s.

PRESS AND BROADCASTING

The legalization of local, independent broadcasting stations in 1976 radically changed the media landscape. Since then the number of newspapers and magazines published has declined, while commercial television and radio channels have mushroomed. The broadcasting sector is dominated by the three state channels of RAI and by three major commercial channels (Canale 5, Italia 1, and Rete 4). The latter three are owned by Fininvest, a multimedia company controlled by Silvio Berlusconi, who built up a virtual monopoly in the private television, advertising, and publishing sectors before becoming prime minister in 1994. The French channels Antenne 2 and TeleMontecarlo compete for viewers in northern and central Italy. About a dozen additional private stations struggle to secure the remaining one-tenth of the national viewership. Italian television has one of the highest numbers of television broadcasts in the EU and produces the largest number of films.

The commercial television sector developed in a legislative vacuum for its first decade after 1976. This had adverse effects for other sectors of the media. Because of its high viewing figures, television drew the major share of advertising revenue away from its habitual market at film theatres and in print media. The effects were especially disastrous for the cinema, but newspapers and magazines also suffered from lack of advertising revenue. As it became increasingly difficult for publishers to operate their newspapers and magazines at a profit, these were gradually taken over by larger industrial and business concerns, often with some compromising of their editorial freedom. In the 1990s legislation to reorganize the broadcasting industry—to prevent the creation of monopolies and to regulate restrictions on the press—proved highly contentious.

Media consolidation

The major national dailies are *Corriere della Sera, La Repubblica, La Stampa*, and *Il Giorno*. Local papers are vital in Italy, underlining the strength of regional identity. Among the newspapers with the largest circulation are the sports titles *La Gazzetta dello Sport* and *Corriere dello Sport*. (R.L.K./M.F.K./P.E.S./Ed.)

HISTORY

Italy in the early Middle Ages

The Roman Empire was an international political system with Italy only a part of the whole, though an important part. When the empire fell, it was replaced in Italy, after the Lombard invasion of 568–569, by a network of smaller political entities. How each of these developed, in parallel with the others, out of the ruins of the Roman world, is one principal theme of this section. The survival and development of the Roman city is another. The urban focus of politics and economic life inherited from the Romans continued and expanded in the early Middle Ages and was the unifying element in the development of Italy's regions.

For the history of the Roman Empire, see the *Macropædia* article GREEK AND ROMAN CIVILIZATIONS, ANCIENT.

THE LATE ROMAN EMPIRE AND THE OSTROGOTHS

Under the military emperors of the late 3rd century, most notably Diocletian (284–305), the political structures of the Roman Empire were put onto a new footing. The army was restructured after the disasters of the previous 50 years; the civil bureaucracy and the ceremonial rituals of imperial rule were considerably developed; and, above all, the tax system was reorganized and enlarged. The fiscal weight of the late Roman Empire was heavy, given the resources of the period: its major support, the land tax, collected by local city governments, took at the minimum one-fifth, and probably one-third, of the agricultural produce. On the other hand, the administration and the army that the tax system paid for reestablished a measure of stability for the empire in the 4th century. Central government was not always stable; there were several periods of civil war in the 4th century, notably in the decade after Diocletian's retirement and in the years around 390. But succession disputes had been a normal part of imperial politics since the Julio-Claudians in the 1st century AD; in general, self-confidence in the 4th-century empire was fairly high. Aggressive emperors such as Valentinian I (364–375) could not have been able to imagine that within a century nearly all of the Western Empire was to be under "barbarian" rule. Nor was this lack of a sense of doom a simple delusion; after all, in the richer Eastern provinces the imperial system held firm for many centuries, in the form of the Byzantine Empire.

Fifth-century political trends. The Germanic invasions of the years after 400 did not, then, strike at an enfeebled political system. But in facing them, ultimately unsuccessfully, Roman emperors and generals found themselves in a steadily weaker position, and much of the coherence of the late Roman state dissolved in the environment of the continuous emergencies of the 5th century. One of the tasks of the historian has to be to assess the extent of the survival of Roman institutions in each of the regions of the West conquered by the Germans, for this varied greatly. It was considerable in the North Africa of the Vandals, for example, as Africa was a rich and stable province and was conquered relatively quickly (429–442); it was more limited in northern Gaul, a less Romanized area to begin with, which experienced 80 years of war and confusion (406–486) before it finally came under the control of the Franks. In Italy, the 4th-century system remained relatively unchanged for a long time. The government of the Western Empire, which was permanently based at Ravenna after 402, became progressively weaker, but it remained substantially intact. Italy, furthermore, was not conquered by a Germanic tribe until the Ostrogothic invasion in 489–493. Although the peninsula faced invasions, such as those of Alaric the Visigoth in 402–410, Italian politics continued throughout the 5th century to be those of the Roman Empire. This meant, in the context of the military crisis of the period, a continual struggle between civil and military leaders, with the emperors themselves more or less as pawns in the middle.

Survival of Roman institutions

The careers of three of these leaders can be taken as examples of 5th-century political trends. Aetius controlled the armies of the West between 429 and his murder in 454; he was the last man to be active in both Italy and Gaul, as a Roman senatorial leader of a largely "barbarian" (German, Hunnic, or both) army. His career was typical of those in the military tradition of Roman politics, and, had his life not been cut short, he might well have become emperor. The makeup of his army was, however, already significantly different from that of Diocletian or Valentinian, and its growing number of non-Roman military detachments tended increasingly to have their own ethnic leaders and to be organized according to their own rules. Ricimer (in power 456–472, by now only in Italy) was a German, not a Roman. He was culturally highly Romanized and, as such, was himself part of a tradition of Romano-Germanic military leadership that went back to the 370s, but he could not, as a German, be emperor, and he made and unmade several emperors in a search for a stable ruler who would not undermine his own power. Significantly, in 456–457 and 465–467 he ruled alone, subordinate only to the Eastern emperor in Constantinople. Odoacer was militarily supreme from 476 to 493, replacing the last ethnic Roman military commander, Orestes (whose son was the child emperor, Romulus Augustulus), in a coup. He pushed Ricimer's politics to its logical conclusion and ruled without an emperor, except, once again, for the nominal recognition of Constantinople as supreme authority. Odoacer, however, did not only call himself *patricius*—local ruler for the Eastern Empire—but also *rex*—king of his Germanic army of Sciri, Rugians, and Heruls. How far he was a military commander of a Roman army and how far a German "tribal" leader was by now impossible to tell. Nonetheless, he, like Ricimer, was an effective defender of Italy against invaders for a long time.

The Ostrogothic kingdom. Odoacer was conquered and killed by Theoderic, king of the Ostrogoths (489–526). The decades of the Ostrogothic kingdom in Italy (489–552) can be seen as the first true period of Germanic rule in the peninsula, for an entire tribe of 100,000 to 200,000 people came with Theoderic. Still, the Ostrogothic kingdom continued to operate inside a largely Roman political system. Like Odoacer, Theoderic courted the Roman aristocracy, both the civil administrators at Ravenna and the great landowners who made up the Senate at Rome. He needed them to run a still largely functioning tax system, which continued, in part, to pay for the army, though the latter was now entirely Ostrogothic. Roman law remained the basis of political and civil life except for the Ostrogoths, who kept themselves apart. They were encouraged in this by Theoderic, who did not want them to become Romanized. Yet such apartheid did not last. Some Romans joined the army; many more Goths became landowners, legally or illegally, and adopted civilian Roman cultural traditions.

The Ostrogothic kingdom

Theoderic's rule has always been seen as the most peaceful and prosperous period of Italian history since Valentinian, and not wrongly; but it was already in ruins a decade after his death. Theoderic himself fell out with an important, traditionalist, senatorial faction, and he executed several senators, including the philosopher-politician Boethius, in 524; the Roman elites looked increasingly to Constantinople as a result. The Goths began to split between factions representing more Roman or more German cultural traditions; when the latter faction murdered Theoderic's daughter and successor Amalasuntha (regent 526–534; queen 534–535), a crisis began that was to end the kingdom.

The end of the Roman world. The Eastern emperors in Constantinople regarded themselves as the legitimate rulers of the West, including Italy, after 476; both Odoacer and, for a time, Theoderic had recognized them, and they had strong links with the Roman Senate. In 533–534 Belisarius, general for the Eastern emperor, Justinian (527–565), conquered Vandal Africa; Amalasuntha's death was the necessary excuse to invade Italy. Belisarius arrived in Sicily in 535, and by 540 he had fought his way north to Ravenna. The Ostrogothic king Witigis (536–540) surrendered to him. The Gothic armies of the north, however, elected new kings, and Totila (541–552), the most successful of them, kept the war going throughout the peninsula until his death in battle. The Gothic wars were a disaster for Italy; almost no region was untouched by them. Together with the subsequent wars of the Lombard settlement (568–605), they mark the end of the Roman world there. In the 550s and the early 560s, however, the Eastern (thenceforth, Byzantine) Empire succeeded in reestablishing its political order in Italy, and in 554 Justinian issued a Pragmatic Sanction setting forth its terms: Italy was made a province of the Byzantine Empire, with its capital still at Ravenna (Sicily, Sardinia, and Corsica, however, were to remain administratively separate), and the Ostrogothic political system was to be dissolved. Indeed, the Ostrogoths virtually vanished as a people from then on; they must have been absorbed into the Roman population or into that of the Lombards.

The Gothic Wars

LOMBARDS AND BYZANTINES

In 568–569 a different Germanic tribe, the Lombards, invaded Italy under their king Alboin (*c.* 560–572). They came from Pannonia (modern Hungary), which had itself been a Roman province. Exactly how Romanized they were is a matter of dispute, but they certainly did not have the political coherence of the Ostrogoths, and they never conquered the whole of Italy. Alboin took the north but was soon murdered, probably with Byzantine connivance. His successor, Cleph (572–574), was murdered as well, and then for a decade (574–584) the Lombards broke up into local duchies, with no king at all. The Byzantines seem to have been partially responsible for this too; at that time they did not have the military capacity to drive the invaders back, and it was easier for them to divide the Lombard leadership and buy some of them into the Byzantine camp. For the rest of the century, even after the reestablishment of Lombard kingship under Authari (584–590) and Agilulf (590–616), nearly as many Lombard leaders seem to have been fighting with the Byzantines as against them. In 584, in the face of Frankish invasions from beyond the Alps, the Lombard dukes met and elected Authari king, ceding him considerable lands; on the basis of these, Agilulf managed to unify the duchies of the north into a single kingdom. But the confusion of the first decades of the Lombard kingdom did not favour the development of a coherent political system, and when the wars stopped, in 605, Italy was divided into several pieces, with boundaries that were in some cases to survive for centuries.

The Lombard kingdom

The largest of these pieces was the Lombard kingdom of northern Italy and Tuscany. By the 620s its capital was at Pavia, which remained the capital of the north until the 11th century; other major centres were Verona, Milan, Turin, Lucca, and Cividale, the capital of the duchy of Friuli. Friuli played an important role as the Italian frontier against the Avars, a powerful military confederation of Central Asian origin that had taken over Pannonia. The two great southern duchies of the Lombards, Spoleto in the central Apennines and Benevento in the mountains and plains of the south, are best treated as independent states; they were not connected to the Lombard kingdom geographically and seem to have developed separately, as territories conquered in the 6th century by Lombard detachments originally in some sense under Byzantine control. They were only part of the same political structure as the north for brief periods, most notably the 660s and the 730s–760s.

Byzantine Italy was nominally a single unit, but it too in reality fell into several separate pieces. Its political centre was Ravenna, which was ruled by a military leader appointed from Constantinople and called exarch from about 590. Exarchs were changed quite frequently, probably because military figures far from the centre of the empire who were too well placed locally might revolt (as happened in 619 and 651) or else turn themselves into autonomous rulers. But the impermanence of the exarchs made it easier for their local subordinates to gain some measure of autonomy. The duke of Naples, the largest city of the south, was effectively independent by the 8th century, and so was the duke of the newly formed lagoon city of Venice. The most important of these local rulers, however, was the pope, the bishop of Rome, for Rome remained the largest city of Italy, and its bishop, in theory the spiritual head of the whole of Latin Christendom, had considerable status. Rome had dukes too, but they did not have the local support the popes had, and they remain shadowy figures. The popes, on the other hand, had a political position that in practice equaled that of the exarchs and lasted a great deal longer. In the far south, Sicily remained administratively separate from Ravenna, as did Sardinia, which followed its own path under increasingly independent "judges" in almost total obscurity until the Pisan and Genoese invasions of the 11th and 12th centuries. The Beneventan Lombards took Apulia (now Puglia) from the Byzantines, except for Otranto at its southern tip, in the late 7th century; southern Calabria remained under Constantinopolitan control and was Greek-speaking by the 10th century.

The Lombard kingdom, 584–774. King Authari ensured the survival of the Lombards, threatened as they were by both the Byzantines and the Franks. The last Frankish invasion was in 590 and probably resulted in the establishment of some sort of Frankish supremacy; the Lombards payed tribute, at least for a time, and sent detachments to fight in the Frankish army as late as the 620s. Agilulf reorganized the kingdom and suppressed several dukes with pretensions to autonomy. He also concluded a treaty with the Byzantines in 605 that established permanent borders with the exarchate, which scarcely changed over the next century (the only major exception being the Lombard conquest of the Ligurian coast in the early 640s). Agilulf also seems to have reorganized the central government, with the help of Roman administrators, and indeed he imitated or reestablished some late Roman and Byzantine court rituals; he did not, however, exact the land tax and must have lived for the most part off his substantial royal estates. Agilulf seems to have been a pagan in his personal religion, though he may have been an Arian Christian; there were certainly many Arians among the Lombards, including most of the kings between 568 and 652. His wife and son were, however, Catholic, and Catholics were common among the Lombards as a whole from at least the 590s as well. Germans had often been Arians in the 5th and 6th centuries (the Ostrogoths were, for example); but the Lombards seem to have been less committed to Arianism than were the Goths or the Vandals, and they abandoned it without documented struggle in the mid-7th century. Although the Lombards do not in any case seem to have been religious fanatics, it may well have been Agilulf, once more, who laid the basis for a peaceful conversion of his people to Catholicism, owing to his careful cultivation of links to Catholic figures such as Pope Gregory I (despite his wars with Rome), or the Irish missionary Columban, who founded the monastery of Bobbio near Pavia in 613.

Religious affiliations

For the political history of the Lombards, scholars rely primarily on one source, Paul the Deacon's *History of the Lombards,* written in the 790s. For the reigns of Agilulf and his predecessors, Paul's information is in part contemporary, for it is based on a lost historical work by Secundus of Non, one of the Romans at Agilulf's court. Secundus' work, however, seems to have ended after 616, and Paul's knowledge—and thus posterity's—becomes much more fragmentary. Paul says little, for example, about Rothari (636–652), except that he was militarily successful (it was he who conquered Liguria) and, most importantly, that he was the first king to set out Lombard custom, in his Edict of 643, a substantial law-code that survives independently. It is evident, however, that the basic institutions of the kingdom were by then fairly stable. Between 616 and 712 the succession was dominated by the Bavarian dynasty, the family of Agilulf's wife, Theodelinda; kings who were not members of this family, such as Rothari and Grimoald of Benevento (662–671), married into it. Grimoald was the only southern duke to aim at the throne of Pavia; like Rothari, he fought the Byzantines and made laws. Male-line Bavarian kings such as Perctarit (661–662, 672–688) and his son Cunipert (680–700) preferred peace and seem to have developed the ceremonial role of the royal court. This contrast may have represented a real political difference, but, if so, it was only a difference of emphasis. The cornerstones of the Lombard political tradition, Agilulf's Romanized court and Rothari's Lombard law, came to be accepted by every king.

The Lombard political succession, like that of the Visigoths in Spain, was dominated by coups, and between 700 and 712 these became particularly savage, resulting in the end of the Bavarian dynasty. Liutprand (712–744) reestablished peace; he is generally regarded as the most successful Lombard king. He issued a series of laws, as a conscious and well-organized updating of Rothari's Edict, which introduced a fair amount of Roman law into the Lombard system. He also waged war on the exarchate and the southern duchies alike. The duchies of Spoleto and Benevento had, as noted, maintained their independence and their separate political traditions. Liutprand conquered the southern duchies in the 730s, setting up his own dukes in both; by his death, Spoleto (though not Benevento) was stably in Pavia's orbit. He also took about half the land controlled by the exarch and occupied Ravenna itself, temporarily, in 743. His attitude to Rome is less clear; he took some papal territory but never threatened the city itself. During Liutprand's reign the Lombard king, for the first time since 568, was militarily dominant in the peninsula. He seems, however, to have still regarded the exarch and the pope as having some right to an independent existence.

The reign of Liutprand

Aistulf (749–756) followed Liutprand's policies to their logical conclusion: he conquered Ravenna in 751, ending the exarchate; he ruled in Spoleto without a duke in 751–756; and in 752 he began to move on Rome, demanding tribute from the pope. But times had changed for the Lombards. In the 740s the popes had become close to the rising Carolingian dynasty in Francia, and in 751 its head, Pepin III, was recognized as king of the Franks by Pope Zacharias (741–752). Faced with Aistulf's attacks, Zacharias' successor, Stephen II (752–757), went to the Franks and sought their military support. In 754, and again in 756, Pepin invaded Italy and defeated Aistulf; he took Ravenna from the Lombard king and gave it directly to the pope, notwithstanding protests both from Byzantium and from the inhabitants of Ravenna itself. This pattern was to persist. Aistulf's successor, Desiderius (757–774), allied himself by marriage with the Franks and kept control of the southern duchies. But when he too threatened Rome in 772–773, the Frankish king, Charlemagne, invaded and this time conquered the Lombard kingdom outright (773–774). Italy became absorbed into the Carolingian lands right down to the border of Benevento, which remained independent.

Popes and exarchs, 590–800. The Byzantine lands in Italy were, of course, in theory only provinces of the empire of Constantinople and to that extent do not have much of an independent political history. Although Ravenna often found itself on the opposite side politically from Constantinople, few exarchs made a permanent impression. The most consistent local political tradition was probably that of the archbishops of Ravenna, who were rich and powerful and, like their counterparts in Rome, had a considerable role in the civil administration.

It was in this context that the popes gradually increased their secular authority. The exarchs did relatively little to defend Rome, which was largely cut off from Ravenna by the Lombard states; the papal city thus had to develop its own political traditions. In the late 6th century, responsibility for feeding the population of Rome and, by the 590s, for defending it from the Lombards (both of Pavia and Spoleto) slowly shifted from the fast-disintegrating Roman Senate to the popes, who themselves still tended to come from senatorial families. Gregory I the Great (590–604), was the most important of these, and, thanks to his own extensive theological writings and collection of letters, his papacy is by far the best documented. In the course of the 7th century, his successors slowly detached themselves from the power of the exarchs, and by

Italy in the year 700 AD.

about 700 they could successfully defy any attempt from Ravenna to remove them. This also meant that they had gained autonomy from the more distant authority of the Byzantine emperor, with whom they were also often in religious disagreement. Pope Martin I could in 653 still be arrested for such disagreement (he died in exile in the East in 655), but not his successors. This autonomy became particularly important in the 730s, when the emperors were iconoclasts (*i.e.*, opposed to religious images, or icons) and the popes were firmly opposed to iconoclasm. The emperor Leo III (717–741) confiscated papal rights in southern Italy and Sicily from Rome for its defiance, but he could not remove the pope. From then on, however, the Byzantine army no longer helped the popes, who were increasingly reliant on their lands in the Campagna around Rome for food and military support. It was in this context that the popes began to look to the Franks for help when they needed it against the Lombards. But the popes were also, in the face of nothing but hostility from Byzantium, beginning to think for the first time in terms of their own practical independence. This came to fruition when the popes gained control over Ravenna itself after 756; by 774, when Charlemagne conquered northern and central Italy, Pope Adrian I (772–795) had extensive territorial designs in the peninsula. These came to nothing, and indeed Adrian and Pope Leo III (795–816) found Charlemagne a far more intrusive patron than the Byzantines had ever been. But the popes kept control of the Roman Campagna, and the Rome-Ravenna link remained intact as well; the Papal States, as reconstituted by the late medieval popes, reproduced almost exactly the boundaries of the former exarchate.

Papal independence from Byzantium

Ethnic identity and government. *Lombard Italy.* The Ostrogothic kingdom used so many Roman governmental institutions that it can most easily be analyzed as if it were part of the late Roman imperial system. Lombard rule marked much more of a break, without doubt. But exactly how much the Lombard states owed to the Roman past and how much to Germanic traditions is still an ongoing debate. The basic notion of the kingdom as a political system was a Germanic concept in large part, for the legitimacy of the king rested on his direct relationship with the free Lombard people in arms—the *exercitales,* or *arimanni,* who formed the basis of the Lombard army. This concept did not leave much room for Romans, who indeed largely disappear from the evidence, even when documents increase again in the 8th century; it is likely that any Romans who wished to remain politically important in the Lombard kingdom had to become "Lombardized." It is even in dispute, for that matter, how many such Romans there were; Paul the Deacon, for instance, claimed that the Roman aristocracy were largely killed in the first generation of the Lombard invasion. But this was certainly an exaggeration, because the Lombards adopted too many customs from the Romans for the latter to have been reduced entirely to subjection. Some Roman aristocratic families must have survived among the Lombards, as is suggested, for example, by the name of a royal protégé and founder of a monastery in Pavia in 714, Senator, son of Albinus.

The Lombard settlement seems to have been largely to the north of the Po River, the area with the majority of Lombard place-names and Germanic-style archaeological finds (mostly from cemetery sites). But even there Lombards must have been in a minority, and they must have been even more so farther south. There were probably few concentrations of Germanic settlers entirely immune to Roman cultural influence. The Lombard language seems to have disappeared by the 8th century, leaving few loanwords in the Italian language. The impression conveyed is of a gradual Romanization of the society and culture of the Lombards within the framework of their continuing political dominance. When the Franks invaded, Lombards and Romans moved together still more as a conquered, by now "Italian," people: the *regnum Langobardorum* of the Lombard period was called the *regnum Italiae* from the 9th century onward.

Romanization of Lombard society and culture

This pattern is reinforced by the evidence of Lombard law. Rothari's Edict and Liutprand's laws look much like the legislation of the Franks and of other Germanic peoples; they deal, for example, with the carefully calculated compensations for various crimes of violence that were designed to replace violent feud or at least to make easier the resolution of feuding. These were certainly ideas foreign to traditional Roman law. When Liutprand in 731 restricted the scope of the judicial duel, for he suspected that it was unjust, he explicitly recognized that it could not be abandoned altogether, for it was part of Lombard custom. Within this Lombard frame, however, the content of law was often in practice heavily Roman. Lombard land law, for example, was almost entirely late Roman, except for the rules for inheritance.

The administrative system of the Lombard state was even more Roman than its laws. This is not very surprising, for Roman models offered far more power to rulers than did any Germanic tradition of government. The Lombards, like other Germanic invaders, took what they could from their new subjects and used Roman administrators where they could find them. Their system, as it is visible in documents from the 8th century, seems to have been more coherent than that of most other Romano-Germanic kingdoms. It was based on a central government in Pavia with numerous permanent administrators (such as the *referendarii,* who organized the writing of royal charters) and legal experts; there is evidence of legal appeals to judges in Pavia, and some of them were settled by the king himself.

Romanization of the administrative system

Locally, government was mostly based on cities, which was another Roman tradition. In the kingdom, cities were governed by either a duke or a *gastald;* the difference seems to have been principally one of status. In the southern duchies, local rulers were all *gastald*s. These officials were in charge of the local law-courts, led the city army, and administered the royal lands in the city's territory. (These three duties more or less exhausted the functions of government in the early Middle Ages.) Such responsibilities were typical everywhere in the post-Roman world; in Lombard Italy, however, the local power of dukes and *gastald*s seems to have maintained a more official character than in, say, Francia, with less movement toward the development of private, or family, power and more royal intervention in local political processes. The Lombard kingdom also differed from Francia in the relatively limited political importance of its bishops and other churchmen; the kings of Pavia used church institutions as an element to bolster their power less than did any other rulers in the West (including the Byzantines in Italy). This may well show that secular institutions were strong enough for kings to rule through them without ecclesiastical help; if so, the reason must have been the survival of a relatively complex social and political life in the cities themselves. Eighth-century documents, particularly for Lucca, show a network of medium-level aristocratic families based in cities, who tended to furnish both counts and bishops for their localities and whose genealogies can sometimes be traced for centuries to come. The stability of city politics was probably the essential foundation for the political coherence of the Lombard kingdom itself.

Byzantine Italy. Byzantine Italy was different from the Lombard lands in obvious and crucial respects. It was not independent; it was not ruled by an incoming, ethnically distinct group; and it gave more political space to the church. Perhaps above all, it still exacted the land tax and thus could afford a salaried army and a far more complex administrative system than the Lombards ever had. But in some respects it had a very similar development. The local power of the army and the constant need for defense led to the formation of a militarized landed aristocracy and, indeed, to a military identity for local free owners at all levels, and thus to social patterns that were not at all unlike those in the Lombard states. For that matter, the foreign origin of many newly landed army leaders (Greek or Armenian) made the ethnic mix in the Byzantine lands almost as visible as in those of their Germanic neighbours. The civilian aristocracy of the Roman Empire vanished; local Roman landowners who wished to survive as political figures had to become militarized and "Byzantinized," at least if they did not attach themselves to the bureaucratic network around the popes and the archbishops of Ravenna. Even the church became increasingly militarized; by the 9th century the bishop and the duke of Naples were sometimes the same person. The dominance of local military aristocracies in ecclesiastical politics is seen most clearly in the civil wars in Rome in the late 760s, the first period of effective papal independence and one in which rival families fought it out for the papal office. Henceforth Roman politics was to take on this internecine character whenever popes were politically independent; the next sequences of violence occurred in the years around 900 and in the early 11th century.

Similarities between the Lombard states and Byzantine Italy

Similarities between Lombard and Byzantine states. The Lombard states and the Byzantine provinces in Italy thus resembled each other more than either did the Roman Empire of the 5th century. The Lombard kings had a far less complex administrative system than had existed before 550, based as it was on royal landowning rather than the complex tax-raising mechanisms of the Roman world. One example of this is that they usually only minted a high-value gold currency rather than the gold, silver, and bronze coins normal under the empire: their state did not need as complex a financial system as the Romans had had. But the complexity of public life could more easily survive in Lombard Italy than farther north in Germanic Europe, owing above all to the vitality of Italian city society: in this sense, the Lombards looked far more Roman than did the Franks, or, still less, the Anglo-Saxons. This city society must have been fairly similar on both sides of the Lombard-Byzantine frontier, in Ravenna as in Pavia, in Rimini or Naples as in Lucca or Verona. And, as the Byzantines developed local military aristocracies resembling those of the Lombards, so the cultural traditions of the two parts of the peninsula tended to move in the same direction. They were never identical, however; major Byzantine cities seem to have been larger than Lombard ones, and the Byzantine political system remained the more complex and articulated of the two to the end.

CAROLINGIAN AND POST-CAROLINGIAN ITALY, 774–962

The kingdom of Italy. The Carolingian kingdom of Italy occupied the northern and central peninsula down to Rome, with the sole exception of the nominally Byzantine duchy of Venice; the exarchate and all the Lombard lands except Benevento (to be dealt with separately) were part of it. Charlemagne called himself "king of the Franks and the Lombards," thus recognizing the separate identity of Italy inside the Carolingian empire. He left the Lombard dukes and *gastald*s in place unless they openly rebelled against him. Indeed, Italy was so much more tightly governed than Francia that to some extent it served as a model for Charlemagne's governmental reforms. However, these reforms were intended for the entire empire, and, in general, the reign of Charlemagne in Italy (774–814) effected the slow integration of the latter into the political world of the Franks. Frankish names for institutions and offices came to replace Italian ones; for example, dukes and *gastald*s became counts, *gasindi* (private military dependents) became *vassi* (vassals), and minor law-court officials were henceforth called *scabini,* as north of the Alps. As in Francia, the church came to be of greater political importance, for the Carolingians in Italy used bishops in their central and local administrations almost as much as they used counts. And, as long as the Carolingian empire remained united, its legislation, with some modifications, was as valid south of the Alps as it was to the north. The Frankish conquest began, then, a period of slow change rather than rupture: certainly, there was less rupture than with the Lombard conquest in 568. Few Franks, in fact, settled in Italy. These were mostly aristocrats, and indeed they made up almost the entire body of Italian counts appointed after about 800; but the Lombard aristocracy remained in the cities, and most bishops continued to be from their numbers—and bishops were steadily gaining in political importance.

Charlemagne's governmental reforms

Carolingian government is better documented than that of the Lombards and seems to have slowly increased in sophistication: Carolingian rule in northern and central Italy (774–887) brought a century of uninterrupted peace, and kings had time to perfect the already systematic ties between Pavia and the increasingly literate city administrations. King's messengers regularly brought royal commands to the cities, and appeals came back to a complex judicial network in Pavia. Locally, legal procedures became standardized and reliable, as surviving documents of court cases show. This does not mean that government or law were equitable or just, and there is plenty of evidence to indicate they were not; but they were at least systematic. This administrative network remained, even after the crisis of royal power in the early 10th century.

For most of the 70 years after 774, the kings of Italy were either children or living in the north of Europe. Charlemagne rarely came to Italy; his son Louis the Pious (814–840) never did. Charlemagne was at least involved with Roman politics, and Leo III crowned him emperor in 800; but this title held little practical significance until the German emperors reestablished it in 962, and Louis was anyway crowned emperor in Aachen (Germany), not in Rome, in 813. Louis' brother Pepin was sub-king of Italy until his death in 810, and he was succeeded by his son Bernard (812–817). Louis, however, replaced Bernard with his own son Lothair (or Lothar; 817–855); Bernard revolted, but he was captured and blinded, and he died in 818. Lothair, like his father and grandfather, was more interested in Frankish politics, particularly during the Frankish civil wars of the 830s. After the Treaty of Verdun in 843, the Carolingian empire began to be divided between the male heirs of the dynasty; West Francia (roughly, modern France), East Francia (roughly, modern Germany), and Italy were the major new kingdoms that emerged. Lothair's son Louis II (844–875) was king-emperor only in Italy. Louis II, whose reign was in many ways the high point of the Carolingian kingdom in the peninsula, was an active interventionist king. He used both the Pavia administration and new legislation to restore royal authority, which had slipped a little during the civil wars in Francia. His laws of 850, in particular, directed against robbery and the abuse of power by the rich, attest to the seriousness of his intent. Louis, and an entourage of powerful bishops and lay aristocrats (notably the Supponids, relatives of his wife Engilberga), reestablished firm royal hegemony in northern Italy in the 850s.

Louis II

The role of Rome. Rome was in practice part of Carolingian Italy, but the popes had a great deal of autonomy and also religious status (Nicholas I [858–867] was particularly influential in Francia). The 9th-century popes were in control of a complex local administrative apparatus and, like their predecessors, played an important role in military defense, particularly against Arab sea raids from North Africa and Sicily (which was conquered by the Arabs in the years 827–902). Leo IV (847–855), in particular, refortified Rome; John VIII (872–882) tried hard to develop military alliances against the Arabs; John X (914–928) eventually succeeded in this, and a coalition of local cities uprooted the Arabs from their stronghold on the sea near Gaeta in 915. The Arabs were a threat to southern Italy, too, particularly after they occupied Bari in 847 during the Beneventan civil war (839–849). Louis II had helped to negotiate an end to that war and was interested in rebuilding Liutprand's southern hegemony. In 866–867 he called up a large army, probably the largest seen in Italy in the entire century, and marched on Bari, which fell to a Frankish, Beneventan, and Byzantine coalition (owing largely to a Byzantine-Slavic naval blockade) in 871. Louis, however, did not leave the south; the Beneventans had to capture him and hold him prisoner for a few days to induce him to return home. This debacle ended Carolingian attempts at hegemony over the entire peninsula; their Ottonian successors were to have no better luck.

The reign of Berengar I. Louis II died in 875, without male heirs. He was succeeded by a series of short-lived uncles and cousins, who came from either France or Germany and stayed in Italy as short a time as possible. But after the fall of the last of these, Charles the Fat (king in Italy 879–887), most of the Carolingian kingdoms turned to non-Carolingian aristocratic families to rule them; in Italy, Berengar I, a female-line Carolingian and also marquis of the still important border area of Friuli, was well placed to be elected as a king with genuine Italian commitments in 888. However, since Carolingians did not have the monopoly over the succession, anyone could claim the kingship; and indeed, Berengar during his long reign (888–924) faced five such rivals, most of them militarily more successful than he was. Berengar was, in fact, not only long-lived but also unpopular; he spent much of the early part of his reign confined in his power base, Friuli. Even when he did not have internal rivals, as in 898–900, he was unlucky; in 899 the Hungarians invaded Italy, destroying Berengar's army and initiating a series of raids that were to last, off and on, until the 950s.

Berengar I's reign was a key period in Italian history. At its beginning, the Italian kingdom was still a powerful and coherent institution, worth fighting civil wars to control. By his death, the relevance of kingship itself was in doubt. This development was partly the result of Berengar's personality, which was unadventurous and militarily unusually inept, but only partly. As the Carolingian political system had settled in, over four generations, local politics had become more stable and inward-looking. Hereditary families had taken over many counties, particularly the big marches of Friuli, Tuscany, and Spoleto. Sometimes local power was balanced between count and bishop, and the kings' capacity to intervene locally increasingly depended on their ability to maintain this balance of power. They usually accomplished this by supporting bishops, conceding more judicial and administrative power to them, particularly after 888. Sometimes, as at Bergamo or Cremona, counts were excluded from inside the city walls altogether. This was occasionally dangerous, for bishops, however loyal, were not royal officials and were more interested in the politics of the city rather than in those of the kingdom; it also represented a clear move toward both the institutionalization of local power autonomous from kings and the fragmentation of that power. In the face of the Hungarian danger, Berengar took this development one step further and localized military defense; after 900 he issued large numbers of grants to private persons, lay and ecclesiastical, of rights to build and fortify castles. His intention was carefully strategic, and his defense in depth was quite effective, but these castles in turn slowly became local centres of personalized military power, and

Institutionalization and fragmentation of local power they gained rights of private justice by the 11th century as well. Carolingian government had always worked better when strengthened by private relationships of a political and military nature; for example, counts relied on their vassals more than on other subordinates to do their bidding, for vassals had sworn personal oaths of loyalty to them. In the castles of the 10th century, personal military bonds became the basis for effective local action. The office of count, too, was to become more and more the basis for private family power, particularly with the appearance in the early 10th century of newer comital families with primarily local roots and fewer national pretensions. Cities remained important administrative centres, but they tended increasingly to be points of reference for the family politics of the military aristocracy rather than bases for royal intervention. These processes had begun well before 888 and were not to be complete until the 11th century, but it is arguable that Berengar's reign marked the turning point. They were crucial for the development of later urban autonomies, culminating in the city communes, but they were disastrous for kings.

Berengar gained 15 years of unopposed rule (905–921) by his cessions of rights and lands after 900. But power was slipping away. Tuscany and Spoleto were semi-autonomous under their marquises; so was the Rome of John X and of Marozia and her son, the *princeps* (prince) Alberic, able and effective rulers between 924 and 954. Hugh of Arles (king 926–947) found the situation irreversible. He could no longer use Carolingian-style procedures, such as new legislation or local administrative intervention, to assert his power. His most typical solution was to overthrow all obvious rivals and replace them with his own relatives, who would in theory be more loyal to him. As a result, he seemed simply violent and high-handed. But the fact is that royal power by now seemed to consist of outside intervention; kings, though still influential and rich, were outsiders to most of Italy. When Hugh faced a coup in 945, his support melted away, and he fell. When Otto I of Germany conquered the Italian kingdom, almost bloodlessly, in 962, his entirely non-Italian power base may simply have seemed to the Italians the logical conclusion of this increasing marginality. The Italian kingdom was to survive as a coherent administrative structure at least until the 1080s, and Frederick Barbarossa even in the 1150s could seek to revive it with some success; but it was by now external to the immediate interests of most of its subjects. After Hugh, no king could have established stable power in the peninsula without a foreign power base and a foreign army.

The marginality of the Italian kingdom

The south, 774–1000. When Charlemagne conquered central and northern Italy, Duke Arichis II of Benevento (758–787) responded by titling himself *princeps* and claiming the legitimist tradition of the Lombards. Lombard princes then ruled in the south for 300 years, until the Norman Conquest. Arichis and his son Grimoald III (787–806) were powerful rulers who held off the Franks, even if Grimoald temporarily had to pay tribute to Charlemagne after an invasion in 787. They controlled the entire southern mainland except for the bay of Naples and the end of the heel and toe of the peninsula, using a governmental system similar to that in the north. But this area is largely barren mountain land and difficult to rule completely; many of the remoter *gastald*s were independent-minded and resentful of Beneventan power. Two of the early 9th-century princes were murdered in aristocratic plots—Grimoald IV in 817 and Sicard in 839; the second of these plots sparked a 10-year civil war that resulted, in 849, in the creation of two rival principalities, based at Benevento and Salerno. The *gastald,* or count, of Capua, Landulf I (815–843), also was interested in independence, and by the end of the century Capua was in effect a third state in the old Beneventan principality.

Move toward local autonomy

Even Naples, though much smaller, was affected by this move toward local autonomy, for the mid-9th century saw the effective secession of nearby Amalfi from Neapolitan control, and the consuls of Gaeta, on the coast toward Rome, were autonomous from the 860s onward. These three cities, like Venice in the Adriatic, were becoming important maritime powers in this century; Salerno was to join them later. The disintegration of the political system of the 8th century was pushed further by the Arabs, who conquered Sicily from the Byzantines after 827 and established bases such as Bari on the coasts of the Italian mainland from the 840s. The Gaetans and the Amalfitans probably owed much of their naval activity and early commercial development to alliances with the Arabs; but others found the Arabs a rather serious danger, notably Bari's neighbours in Apulia and the great monasteries inland from Capua—Monte Cassino and San Vincenzo al Volturno, which were sacked in 883 and 881, respectively.

It was this confused world that Louis II wished to dominate in his great expedition of 867–871, but he failed. More successful was the Byzantine emperor Basil I (867–886), who followed up his blockade of Bari with a set of campaigns that aimed at taking the whole southern mainland from the Lombard princes. Shortly after his death, the latter were pushed out of the plains of Apulia, and by 900 only parts of the Capua-Salerno plain and of the south-central Apennines remained Lombard. In that year the count of Capua, Atenulf I, conquered Benevento, and the Lombard-Byzantine border stabilized. Capua-Benevento maintained a certain cohesion under a single dynasty until the 980s, its most notable prince being Pandulf I "Ironhead" (961–981).

After the departure of the Arabs (except from Sicily) and the straightening out of the political boundaries, the south was much more peaceful in the 10th century than it had been in the 9th. It was dominated by the Byzantines, whose local ruler, or catepan, was head of an administrative and fiscal system that seems to have been more complex and stable than that of the exarchs had been. The Byzantines were by now culturally entirely Greek, and Calabria was, as already noted, Greek-speaking; in Apulia, however, the Latin-speaking Lombards were in the majority, and the Byzantines had to rule through them. They managed this effectively until a series of urban uprisings in 1009–18 brought more autonomy for the Apulian cities—as well as the first Norman mercenaries.

Byzantine predominance in the south

The Lombard states and the independent coastal cities were much weaker. They recognized some sort of Byzantine hegemony, except for the brief periods when the Ottonian emperors sent armies from the north. Their internal structures were less coherent than those of the catepanate. During the 10th century, castles were built everywhere in southern Italy, just as in the Po plain; in the south, however (including the papal territories and the march of Spoleto), their social effect was in many areas more considerable than in the north, because all the scattered population living in the territory of a castle tended to move, or be moved, inside its walls. This process, called in Italian *incastellamento,* created a network of fortified hilltop settlements that often still survive. It could be directed and controlled by the state, as in the Byzantine lands, but in Lombard areas it was effected by private landowners, which greatly extended their local control. The Lombard princes were incapable of dominating this steady political localization, particularly in the mountains. Their attention became concentrated on the richer plains between Gaeta and Salerno. Unfortunately, in this small area there were by now six independent states—Gaeta, Capua, Benevento (when it regained independence in the 980s), Naples, Amalfi, and Salerno. These spent a great deal of time fighting each other after Pandulf I's death in 981, and the Normans in the next century had little difficulty conquering them. The only local success story was international trade, which benefited all the coastal cities (Amalfi being the best known); their fleets had good relationships with Arabs, Byzantines, and Latin Christians and conveyed goods between all three. They dominated long-distance commerce in the western Mediterranean until the rise of the more militarily aggressive cities in the north—Genoa and Pisa—in the 11th century.

LITERATURE AND ART

The early Middle Ages saw relatively few complex literary works; the elaborate educational system of the Roman Empire depended on a level of aristocratic wealth and a style of civilian culture that did not outlast the Gothic

The decline of literary works

wars, and the ecclesiastical educational traditions that succeeded it were not well rooted in Italy outside Rome until the 9th century. Italy's and antiquity's last great philosopher, Boethius (d. 524), had no successors, nor did Pope Gregory the Great (d. 604) in the field of theology. Hagiography, an important early medieval genre in Francia, became almost unknown in Latin Italy after Gregory the Great's *Dialogues.* The writing of history, too, was only rarely practiced in this period: Paul the Deacon's *History of the Lombards,* dating from the 790s, is far shorter than Gregory of Tours' history of the Franks or Bede's of the English, and it had few parallels except for episcopal histories in Rome, Ravenna, and Naples. Nor did the *Rule of St. Benedict,* written by Benedict of Nursia (d. *c.* 547) for his monastery of Monte Cassino, have immediate successors, and indeed as yet it had relatively little effect on Italian culture: 8th-century monasteries did follow it, but the *Rule* owes its international importance to the Anglo-Saxons and to the patronage of Louis the Pious' court in Francia.

Italy did not lose all its cultural traditions, for it developed new ones around the emerging centres of political power of the early Middle Ages. Rome kept up a level of intellectual life owing largely to its links with the Greek culture of the East; it experienced a flowering of new writing in the 9th century around international figures such as Anastasius the Librarian (d. *c.* 878), who had contacts with both Constantinople and the courts of the Frankish kings. Pavia, for its part, developed a largely secular court culture; Paul the Deacon, who was a poet and orator as well as a historian, was partially trained there, and so, later, was Liutprand of Cremona (d. *c.* 972), whose *Antapodosis* is a florid but highly literate satire on the kings of the first half of the 10th century. Charlemagne's court drew Italian intellectuals to it and away from the peninsula; but Carolingian patronage returned to the cities of northern Italy in the mid-9th century, and systematic literary education began to develop in several of them. Tenth-century writers included not only Liutprand but also Atto of Vercelli (d. 961), whose denunciations of contemporary society are written in a Latin so difficult that few have ever understood it. The major intellectual activity in Italy was, however, law. The lawyers at Pavia were already a numerous group in the 9th century; in the 10th century they undertook a large-scale compilation of Lombard law and its Carolingian updatings, usually called the *Liber Papiensis.* This text was the source for 11th-century glosses and expositions and juristic arguments over legal theory that led directly to the 12th-century revival of Roman law at Bologna. The study of law in the Lombard and Carolingian capital may turn out to have been early medieval Italy's major contribution to the development of intellectual life in Europe.

The visual arts

The visual arts showed a more obvious continuity. The architects of Ravenna's monumental mosaic churches and secular buildings from the Ostrogothic kingdom and the years following the Byzantine reconquest developed new styles, but did so as an expansion of late Roman ideologies of public building along Byzantine lines. In Ravenna the great period had ended by 700; in Rome, however, the same tradition continued, if at a reduced level, throughout the early Middle Ages. Sixth-century popes were builders, and 7th-century ones, though less ambitious, were at least rebuilders; from Adrian I onward there was an intense revival reaching its height with large, richly decorated constructions such as the church of Santa Prassede built by Paschal I (817–824). Rome's surviving early medieval buildings are mostly churches, which is not surprising given its rulers; here as elsewhere, however, one must reckon with secular buildings that have not survived and, of course, with a continuous occupation and reuse of the huge array of classical monuments.

In Lombard Italy, building on a monumental scale continued as well, notably in the royal palaces at Pavia and at Monza outside Milan (these do not survive, but Paul the Deacon described parts of the latter). This type of monumental architecture was perhaps associated with a fairly strong tradition of decorative figured stonework, with central European analogues, that survives best at Cividale. One is struck, however, by the small scale of what has been excavated or otherwise studied in the north, as, for instance, the urban monastery of San Salvatore (shortly thereafter renamed Santa Giulia) at Brescia, set up by King Desiderius about 760, or the late 8th-century chapel at Cividale del Friuli, or the tiny frescoed church of Santa Maria at Castelseprio, which may date from the early 9th century. It may be that the Lombards, including their kings, had lost the rhetoric of size that the Romans had had (and that the early medieval Romans kept). The late Roman tradition that survived best was an emphasis on internal decoration, and Italy had many separate schools of fresco painters (as well as, more rarely, mosaicists) by the 9th century. Ninth-century buildings could, however, be large, as was the case with the recently excavated monastic buildings of San Vincenzo al Volturno on the Benevento-Spoleto border, sumptuously frescoed in both northern and southern Italian artistic styles during the first half of the century. Building techniques declined in sophistication in the early medieval period, and older materials were frequently reused. However, good-quality stone and brick in a Roman tradition seem to have continued to be cut and made; it is likely that there were far fewer builders than during the empire but that they continued ancient traditions in major cities. A price book for northern Italian builders from the early 8th century shows that they could make sophisticated private housing. Urban excavations now reveal, however, that more buildings than under the empire were constructed in wood.

Internal decoration

ECONOMY AND SOCIETY

Socioeconomic developments in the countryside. Early medieval Italy was an overwhelmingly agrarian society, as it had been before and as it was to be for long after. Wealth thus derived above all from the ownership of landed estates. Estates were exploited by subsistence tenants on a standard medieval pattern. The slave plantations of 1st-century central Italy had long disappeared, and the word *servus* now usually just meant a tenant without public rights as a free man; the remaining slaves on the land were mostly skilled specialists. Free and servile tenants essentially paid rent, in money or kind, to the estate-centres (*curtes*) of their landlords. For the late 8th and 9th centuries, at least in northern Italy and Tuscany, there is evidence of more organized estates, which were the equivalent of the manors of England and the *villae* of 9th-century northern France; here tenants also did unpaid labour service on the lord's demesne, an area of land all of whose produce went to the lord. These estates, mostly royal or ecclesiastical, could be huge, as were, for example, those of Bobbio and Santa Giulia at Brescia, whose estate records survive. They produced a sizable agricultural surplus, which could easily have been sold in the cities (Santa Giulia, at least, had its own merchants). Not all estates, however, were organized this tightly; elsewhere demesnes, though common, tended to be smaller and less economically central; and in the south of Italy they were always rare.

Organized estates

In the 10th century, Italian landowners increasingly took money rents rather than crops from at least their free tenants, as is known from their surviving written contracts (*libelli*). Money rents were more flexible and could better survive the fragmentation of property between coheirs or its alienation in bits to others, both practices being very common in Italy. But it should be stressed that the very possibility that tenants could pay in coin also makes it clear that there was by now a fair amount of small-scale commercial exchange in the countryside; indeed, the new castles of the 10th century, which were themselves estate centres, typically had markets. In the 10th century, too, more and more servile tenants gained their freedom, whether legally (by formal manumission) or illegally: a law of Otto III in the 990s that intended to restrict the rights of "slaves gasping for freedom" can have had little effect. On the other hand, by 1000, with the development of private judicial powers by landlords over tenants, there were new methods of rural coercion that did not depend on the powers masters had had to punish the unfree, and these methods could be made to apply also to free peasants.

Subsistence cultivation. Italian agriculture was organized for subsistence first; growing crops for sale was rare in the early Middle Ages. Thus, rents in kind tended to reflect what peasants grew for themselves. One finds standard Mediterranean crops such as grain (rye in northern Italy, wheat elsewhere) and wine in 9th-century rent lists; olive oil was common in the centre-south but rare in the north (as it is today), except in specialist farms on the Italian lakes. Early medieval Italy was far more forested than it is today, and peasants seem to have depended substantially on woodland gathering to supplement their diet. Italian peasants probably ate a fair amount of meat, too, more than they were to eat in later centuries. Meat was, however, becoming a sign of an aristocratic life-style by the end of the early Middle Ages; Liutprand of Cremona looked down on the Byzantine emperor Nicephorus II Phocas (963–969) for eating vegetables. Specialist stock-raising was still rare; sheep, cows, and pigs were raised by subsistence cultivators. As a result, cloth and leather were probably not yet made by specialists either, except for luxury goods made by urban craftsmen with an aristocratic clientele. Large-scale urban cloth-working, a central part of high medieval Italian life, still lay in the future. The clearest exception to this was perhaps the linen produced in 10th-century Naples.

Not all subsistence cultivators were tenants; there were many free peasant owners in early medieval Italy. How many of them were descended from small Roman proprietors, how many from Roman tenants who had seized their chance in the confusions of the 6th century, and how many from the rank and file of the Lombard army is unclear. Ethnic Lombards must have been in a small minority, but by the 8th century nearly all landowners in the Italian kingdom professed Lombard law. Most landowning in the 8th and 9th centuries was highly fragmented, with even great landlords owning hundreds or thousands of small parcels of land, scattered among those of other owners, whether aristocratic, peasant, or ecclesiastical. Such a pattern gave a certain independence to village life, where small local owners may often have been quite influential (great lords more often lived in cities, farther away from direct participation in local society). Village communities were, however, usually still informal bodies with little of the coherence they were to gain from the 12th century onward.

Land-holding patterns

The growing power of the aristocracy. The existence of this stratum of free smallholders gave a certain reality to the Lombard, and indeed Frankish, constitutional tradition that based royal power on the nation of free warriors at arms. This tradition was, however, gravely challenged by the rise of the aristocracy. Already in the Lombard period the aristocracy was in practice politically dominant, and probably always had been, in patterns unbroken from the Gothic and Roman period. Yet the 8th-century aristocracy does not seem to have been as wealthy as either its Roman predecessors or its Carolingian and post-Carolingian successors, and this may imply a relative independence for the free peasantry. Under Charlemagne and his descendants, this slowly changed. Incoming Frankish nobles acquired large lands, and churches dramatically increased their landowning. That these developments were often at the expense of the poor is shown by a number of 9th-century court cases, in which peasants claim their land, or sometimes their freedom, usually without success: in some of these cases peasants were clearly in the right. Kings themselves confirmed this, for in the 9th century they worried greatly that the oppressions of the poor would lessen the latter's participation in the public obligations of all free men—*i.e.,* army service, attendance at court, and road and bridge building—and they made laws against such exploitations. The laws were futile, however, and aristocratic landowning and political dominance continued to grow.

In the 10th century, with the breakdown in royal power, these tendencies developed further. In the countryside, castles became the centres of de facto political power that great landowners exercised over their still surviving free neighbours. A new, highly militarized, small nobility began to emerge, based on these castles: their ancestors had been of mixed origins—vassals of counts, local diocesan landowners, or even rising free peasants—but they now held, as a group, a virtual monopoly over armed force; indeed, in the sources they are frequently just called *milites* (soldiers). Counts, where they kept their own power, did so only as leaders of private armies of these *milites,* who, though still their vassals, were now much more autonomous. Churches, to keep control over their extensive lands, had to give much of it out in lease or fief to such military families, and only the strongest churchmen, such as the archbishop of Milan, managed to keep any real power over their new military dependents. This new castle-holding stratum was to become the basic aristocratic class of the 11th to 13th centuries, with only a few of them aspiring to the official titles of count or viscount. Such a tendency was, in fact, common throughout Europe; in Italy the chief difference was that *milites* were never quite as dominant as elsewhere, for cities remained powerful political and military foci, and peasant owners continued to exist in the countryside. The major exception to this was probably the south of Italy, where the new pattern of fortified settlements kept the peasantry inside a more rigid political framework than existed in the more scattered villages of the north. Even inside such a framework of political control, however, some of these fortified villages achieved a new sort of prosperity, for artisanal activity could develop in them, and merchants would come there too.

The new nobility

Socioeconomic developments in the city. Most Roman cities survived into the early Middle Ages as political and economic centres (the majority of the failures were in the Apennines and, to a lesser extent, on the coast). Their function as political centres has already been discussed; there is more dispute about their economic role, however. They must certainly have looked dilapidated, with their Roman monumental centres acting as quarries for rebuilding elsewhere; early medieval public buildings were, as noted, smaller and also probably fewer in number—the cathedral and the local royal palace being the most important ones by far. Archaeologists in cities such as Brescia or Verona have found a much less dense settlement network inside the walls of the early medieval city than in the preceding or later ones, with lower buildings, more courtyards, many more open spaces used for agriculture, and, often, a trend to building in wood. But Roman street plans were maintained in both of these cities and in several others. It is likely that many cities maintained an urban economic identity, with some commercial and artisanal specialization (at least in luxury goods). Lucca's documents in the 8th century show, among others, gold workers, cauldron makers, physicians, and builders, and such figures also appear in texts for Milan and other cities in the 9th century. Essentially, this kind of artisanal activity drew its force from the fact that the city was the residence not only of bishops, dukes, counts, and administrative officials but also of a high proportion of the local aristocracy. The local political interests of the latter can be seen in a wave of competitive church building in the 8th and early 9th centuries; dozens of (probably very small) churches existed in each major centre by 900.

Commerce was undoubtedly far weaker in the early Middle Ages than under the Roman Empire. Archaeology shows it very clearly: the large number of African amphorae and fine ceramics found on every late Roman site in peninsular Italy decreases sharply in the 5th century, and these artifacts vanish in the 6th century. Only from the 8th century onward is there evidence again of pottery exchange networks, but exclusively on the level of the city territory and, as far as is yet known, only around some cities—notably Rome, which remained the largest city in Italy, though it was only a fraction of its former size. City-country exchange networks were probably relatively weak in the 7th and 8th centuries, though they never altogether disappeared. In the 9th century and onward, however, consistent documentary evidence of urban markets shows that these networks were developing again.

Pottery exchange networks

Classical cities had mostly not been major centres of international commerce; or, at least, such commerce was less important as a reason for their existence than the fact that major landowners lived in them. The sharp decline of this commerce in the early Middle Ages was thus not in itself a threat to city life. But its slow revival from about 750 onward did help these cities, for they were still at the nodes of surviving Roman river and road networks that, with few changes, were to become the commercial routes of the high Middle Ages. In the early 8th century King Liutprand issued a text that regulated the salt trade from the Venetian lagoon up the Po River. In the following century this trade developed and increasingly came into the hands of local rather than Venetian merchants; Cremona, among other cities, had become a major mercantile centre by the late 10th century, and not, by then, only for salt—the Venetians, on their way to Pavia, brought, among other wares, spices, ivory, and Byzantine cloth. Venice itself was the focus of this international trade by the 9th century; the will of its duke Giustiniano Parteciaco (also spelled Partecipazio), dating from 829, includes the first reference in medieval history to capital investment, in ships and their goods. By the end of the 10th century the Venetians dominated the trade of the Adriatic Sea and controlled much of its eastern coast. Inland, however, the spin-off of both international and local commerce was bringing a new and visible prosperity to many cities by the 10th century, including Cremona, Pavia—the old political capital still automatically visited by many traders from Venice—the southern cities, and, above all, Milan, which was fast becoming the major economic centre of the Italian kingdom. The northern Italian trade routes, along the Adriatic coasts, up the Po River, and across the Alps, were coming to rival the older routes around the western coast of Italy, via Amalfi and Gaeta (or, later, Pisa and Genoa) and up the Rhône River. Both routes were to develop dramatically in the following centuries. But they did not in themselves create urban life in Italy; that was done by the local aristocracy. The fact that Italian cities remained dominated by landed aristocrats was to condition much of their future history. (C.J.W.)

Italy, 962–1300

ITALY UNDER THE SAXON EMPERORS

In the second half of the 10th century, Italy began a slow recovery from the turmoils of late Carolingian Europe. During the previous century, the Po Valley had been exposed to Magyar raiders. Sardinia, Corsica, and Sicily had fallen to the Muslims; even Rome had felt their threat. In the north, the Lombard kingdom was little more than a collection of great lordships vying with one another for the Carolingian inheritance. In the south, the peninsula was shared by the remnants of the Byzantine and Lombard states and by local powers. The 10th-century papacy had fallen into the hands of various Roman aristocratic factions. But already there were signs of revival. Genoa, Pisa, and Venice were joining other cities in developing local and international trade. In Germany, the last of the East Frankish Carolingians had died, and, in 911, Conrad of Franconia became king, to be succeeded in 919 by the energetic Henry the Fowler, duke of Saxony and founder of the Saxon dynasty of German emperors. In France the Carolingians yielded to the Capetians before the century was out. In the monasteries of Burgundy and Lorraine a new spirit of religious reform arose, which reached outward to the whole of Latin Europe and soon influenced the rich monastic traditions of Italy.

The Ottonian system. In the midst of these favourable signs, the Italian political landscape offered little ground for optimism. The only hope for stability and eventual unity lay with the contenders for the former Carolingian kingdom of Italy. Hugh of Provence cast ambitious eyes across the mountains to the Po Valley; he aimed to pull together the fragments of Lotharingia (Lorraine). But at his death in 947 his son Lothair and later his son's widow, Adelaide of Burgundy, faced strong opposition from Berengario, Marchese d'Ivrea e di Gisla, who assumed the crown of Italy as Berengar II. Adelaide summoned the German king, Otto I (936–973), son of Henry the Fowler, to her aid. Although much involved in affairs in Germany, he came to Italy in 951 and married Adelaide, but he returned quickly to Germany to deal with a rebellion by Liudolf, duke of Swabia, his son from an earlier marriage. Moreover, events in Germany forced him into a confrontation with the Magyars in 955 at the Lechfeld, where his decisive victory ended their attacks on German lands.

Intervention from Germany

At the request of Pope John XII (955–964), Otto returned to Italy, where, in 962, he realized his dream of securing the imperial crown. The coronation of Otto as emperor was, like that of Charlemagne, a recognition of a political reality. Otto was the leading figure among all European rulers of his day. He was a great military victor and a champion of order. He also had built a close alliance with the German bishops. The imperial title, which had dwindled into a virtually worthless symbol, once again legitimated effective political power.

After his coronation, Otto proceeded to consolidate his power by moving against Berengar II, the enemy of his wife's family. Pope John XII, recognizing the emperor's intention of exerting imperial supremacy over the papacy, began to fear for his own future. His activities provoked Otto to move against him. At a Roman synod in December 963 the assembled bishops, mostly loyal supporters of Otto from northern Italy, deposed John and replaced him with Pope Leo VIII (963–965). Otto's decisive action paved the way for his mastery of the kingdom of Italy.

Within two years Berengar was captured. The papacy entered a turbulent decade that ended with the election of Benedict VII (974–983). Otto built his rule on the foundation provided by bishops loyal to the empire; these bishops, many of German origin, owed their promotion to Otto himself. He also relied upon the support of such powerful figures as the marchese of Tuscany and the duke of Spoleto. He pressed his imperial claims with the Byzantines even as he aggressively supported the Latinization of the southern Italian hierarchy (*i.e.,* subjection to the jurisdiction of Rome rather than Constantinople). The chief fruit of his policy in southern Italy was a marriage agreement by which his son, Otto II, became the husband of the Byzantine princess Theophano. Otto had laid the foundation for strong imperial rule in Italy, but he lacked the means to bring it to fruition. Nonetheless, fragile as his foundation in Italy was, it represented a move away from the anarchy of the previous age toward a new era of prosperity and hope for the future.

The yoking of Italy to imperial policy by Otto II (973–983) was an inevitable result of the achievements of his father. One should not, however, view his efforts as a desertion of Germany in quest of the glories of ancient Rome; rather, the policy of the German monarchy, while grounded partly in the idealization of the ancient Roman Empire, had its fullest significance in the vision of Europe derived from the pragmatic realities of the Carolingian age. The strength of its acceptance is already evident in the transfer of power on the death of Otto II in 983 to his son, Otto III (983–1002), a mere child. While the succession did arouse a conflict over the regency in Germany, it was not seriously challenged. The brilliant Gerbert of Aurillac, former abbot of Bobbio and later Pope Sylvester II (999–1003), served as principal adviser and tutor of the young king, whose mother, Theophano, controlled the regency until her death in 991. Otto's grandmother, Adelaide, still an indomitable figure, served as regent until he undertook to rule for himself in 994. Despite his youth, Otto was both able and vigorous. His Italian policy was a continuation of that of his father and grandfather but put in much more explicit terms.

Otto III's Italian policies

Many scholars have argued that Otto III's Byzantine connections shaped his conception of imperial rule. Some have suggested that his ideas were anachronistic; others that he failed to follow the path dictated by the national interests of Germany and Italy. But Otto, who had been schooled in a hard and practical court, aimed in his Italian policies at creating an enduring unity in imperial administration under the imperial chancellor. When his seal employed the style, "Renovatio imperii Romanorum" ("Renewal of the empire of the Romans"), the image invoked was not so much that of Roman antiquity as of the empire of Charlemagne. The renewal referred to a new commitment to the Carolingian design for Europe, viewed from the vantage point of the 10th century. Otto's imperial coronation in 996 by Pope Gregory V (996–999), his own nominee, was reminiscent of that of his grandfather in that he did not hesitate to intervene in Roman affairs. When the Romans drove out Gregory and thought to placate Otto by the election of his former Greek tutor Johannes Philagathus as Pope John XVI (997–998), the emperor returned and in 998 exacted a terrible price from all. He also secured the election of Gerbert of Aurillac as Sylvester II. He did not, however, subscribe to the view of the papal position found in the Donation of Constantine. He rejected this forgery, which purported to list the rights and properties conferred on Pope Sylvester I. Otto supported the claims of the Italian bishops against the lesser aristocracy, who were attempting to make their lands, held from the church, virtually hereditary. For him, as for his predecessors, support of the bishops was a key element in establishing royal control over the cities of central and northern Italy.

Otto III died on Jan. 23, 1002. His body was quickly taken to Aachen (Germany) and laid to rest beside Charlemagne. The German princes elected the duke of Bavaria, who became Henry II (1002–24), the last emperor of the Saxon dynasty. Notwithstanding reassurances to his German supporters of his commitment to effective rule in Germany, Henry's view of his imperial role differed little from that of his Ottonian predecessors. In Italy, he supported the bishops and opposed Arduin of Ivrea, who had seized power after the death of Otto III. It was not, however, until 1013 that Henry was free to come to Italy. After his coronation by Pope Benedict VIII (1012–24) in 1014, he returned to Germany, leaving the bishops the task of disposing of Arduin. In 1021, Henry returned to Italy once more but was unable to extend imperial rule in the south beyond the Lombard principalities of Benevento and Capua.

Social and economic developments. The 10th and early 11th centuries witnessed significant changes in the social and economic life of all parts of Italy. As noted earlier, the upheavals of the 10th century had vastly increased the need for security, leading in the countryside to the fortification of villages. While this process provided security for the peasants, it also strengthened the control over them by both lay and ecclesiastical lords. The conditions of the Ottonian period tended to consolidate this arrangement, given the emperors' reliance on the lay and ecclesiastical aristocracy. The number of great noble families grew rapidly as a direct result of imperial action. These families, often from north of the Alps, were part of the effort to subject Italy more directly to imperial authority. At the same time, however, increases in population, the growth of the cities, and the development of a landed class of knights and lesser nobles (vavasors), began to undermine the Ottonian system based on the support of the bishops and the great marcher lords. The entry of these new social groups into the quest for land created competition not merely between clergy and laity but also within these groups; indeed the interests of clergy and laity were often interconnected.

Vavasors

THE REFORM MOVEMENT AND THE SALIAN EMPERORS

Profound dissatisfactions with the pervasive violence, rapacity, and greed of the age, combined with concerns particularly among the monks about their own vulnerability and that of the poor and weak, fueled a movement for monastic reform. Some early monastic reformers identified their cause with that of the Ottonians. St. Romuald of Camaldoli, for example, worked actively in support of the missionary program of Otto III. The empire represented order and stability, ideals that appealed to those in the monastic profession. But it was also becoming apparent to some that the imperial order impeded efforts to blunt the impact of competition for rights and domains. The reign of Conrad II (1024–39), the first emperor of the Salian dynasty, permitted and even encouraged the development of such competition. Conrad championed the cause of the vavasors, who wanted their lands to be hereditary, at the expense of that of the bishops. At the same time, he supported the interests of the lay aristocracy. Although there is no indication that he intended any permanent change in imperial relations with the bishops (his ties to the papacy were close enough), his actions certainly served to alert Italian ecclesiastical circles to the possibility. But Conrad's influence was not sufficient to cause any major adjustment in relations between the bishops and the empire.

Conrad was succeeded by his son, Henry III (1039–56). Henry was energetic, strong-willed, and devout. He was no innovator, but his attachment to the church served to reduce tensions from the level reached under the rule of his father. Indeed, he resumed the close relations between the crown and the monastic reformers that had characterized the reign of Otto III and Henry II. His Italian policy bears striking resemblance to that of Charlemagne and Otto I. But he lived in different times. His efforts to render decisions that would settle differences among the factions disputing the archbishopric of Milan and his intervention in papal affairs in Rome placed him in the Ottonian tradition. He supported reform, and reform in turn supported the empire. This point has been too often overlooked by historians who have portrayed his actions as inimical to the interests of the empire. This view has been especially influential in discussions of his interventions in papal elections. By emphasizing Henry's piety and his attachment to reform, some historians have de-emphasized the political aspects of his policy. Actually, there was a concurrence between his goals and the desires of the reformers. When Henry arrived in Rome in 1046, he found the papacy in disarray. In the continuing competition among leading Roman families for control of the papacy, the Tusculan faction had elected Benedict IX (1032–44), who was driven from the city and replaced by the candidate of the Crescentians, Sylvester III (1045). Benedict regained the papacy in 1045 but made way for the election of John Gratian as Gregory VI (1045–46), a strong supporter of reform. Henry was therefore confronted by an uncertain situation just when he was seeking imperial coronation. The synods of Sutri and Rome resolved the difficulty by deposing the three previous claimants. At the behest of Henry, the bishop of Bamberg was elected as Clement II (1046–47). The new pope immediately proceeded to Henry's coronation on Christmas Day, 1046. The Carolingian precedent—Charlemagne's coronation also took place on Christmas Day—could hardly have been lost on his audience.

Henry III took Gregory VI back to Germany with him, aiming in this way to prevent a resurgence of internal conflict in Rome. But death soon overtook both Clement and his German successor, Damasus II (1048). Again Henry intervened and secured the election of Bruno of Toul, who took the name Leo IX (1049–54). Leo combined strong attachment to the imperial cause with dedication to the cause of reform. Profoundly influenced by the monastic centres of reform in Germany and Burgundy, he turned especially to them for collaborators in the work of rebuilding the battered prestige of the Roman church. He brought to Rome men like the theologian Frederick of Lorraine, Hugo Candidus, and Humbert of Moyenmoutier, who became cardinal-bishop of Silva Candida (*i.e.*, a bishop of the suburbicarian dioceses). The deacon Hildebrand, who had accompanied Gregory VI to Germany as his secretary, also returned to Rome and joined the papal entourage. Under Leo's leadership, the ancient body of cardinals was transformed into an effective instrument for administration of the church and promotion of reform. Leo held synods in northern Europe and Italy aimed at stirring local commitment for the program of the reformers. That program was chiefly directed at freeing churches from lay control, especially by the appointment of unworthy candidates to ecclesiastical office through simony (*i.e.*, the practice of buying church offices), and at forbidding the pervasive practice of clerical marriage and concubinage, which threatened the substance of the church. Leo's efforts must be viewed against the background of the monastic reform movement, which had already achieved considerable success in regaining control of monastic properties and preventing their further alienation not merely at the hands of the laity but also at those of other ecclesiastics.

The reform program

Although couched in moral terms, the program of the reformers served eminently practical ends.

The papacy and the Normans. Leo remained committed to the imperial ideal. His opposition to the Normans, who had arrived in southern Italy in the early 11th century (see below), arose not merely from his concern for the threat they posed to Rome and the papacy but also from his view that they threatened the interests of the German emperor and relations between East and West. The continued expansion of the Normans in southern Italy and their aggressive assertion of titles—William de Hauteville, (William Iron Arm), for example, assumed the title of count of Apulia—influenced Leo to forge an alliance of papal, imperial, and Byzantine forces. With himself in the company of imperial troops but without awaiting the arrival of promised help from the Byzantines, he met the Normans at Civitate on June 16, 1053. The ensuing defeat was a deep humiliation for Leo, though the Normans treated him with respect. The forced peace profoundly disturbed the balance he had sought in Italy. The future of his policy also received a serious setback from the conflict that arose in Constantinople between his legates, Humbert of Silva Candida, Frederick of Lorraine, and Peter, archbishop of Amalfi, and the Eastern patriarch, Michael Cerularius. The reasons for this conflict have been variously described, but it was at least in part due to the clash between the papal policies of Latinization of the churches in southern Italy and the claims of Constantinople to jurisdiction in that region. The mutual excommunications launched by the legates and the patriarch in 1054, after the death of Leo, have often been viewed as the beginning of the schism between the Eastern and Western churches; however, that view probably overstates the significance of these events. More particularly, the breach with Constantinople closed the door to the approach taken by Leo IX and led to a major shift in papal policy in favour of the Normans.

Battle at Civitate

Leo's successor, Pope Victor II (1055–57), formerly the bishop of Eichstätt (Bavaria) and the fourth pope chosen under the aegis of Henry III, tried to follow in the footsteps of his predecessor. But the death of Henry III (1056) and the failure of Leo's policies in southern Italy limited his role. The Normans continued to strengthen their position in Italy. Having arrived in southern Italy in small groups just prior to 1020, perhaps in part as petitioners at the court of Pope Benedict VIII and in part as pilgrims returning from the Holy Land, the Normans joined Lombard rebels in southern Italy in their effort to throw off Byzantine rule in Bari. Although this proved a failure, Norman mercenaries continued to enlist in the armies of various southern Italian rulers. In 1030 Sergius, duke of Naples, granted the county of Aversa to the Norman Rainulf in return for his support against Pandulf of Capua. Rainulf was able to add Gaeta to his holdings, and his nephew, Count Richard, who had succeeded to Aversa in 1047, added the principality of Capua. The next wave of Normans, led by the sons of a lesser Norman landholder, Tancred of Hauteville, undertook a full-scale effort to conquer the south. Robert Guiscard, Tancred's fourth son, assumed a commanding role in southern Italian affairs. Victor II thus was constrained not merely by the failed mission to Byzantium but also by the threat from the south. Moreover, on Henry III's death, the empire came to his six-year-old son, Henry IV (1056–1106), with his mother, Agnes of Poitou, as the regent. Although the succession to the throne was not in doubt, the inevitable intrigues surrounding the regency deprived the papacy of imperial support. When Victor died in 1057, a party of the reformers moved to take advantage of this vacuum. It acted quickly to elect Frederick of Lorraine as pope, under the name Stephen IX (or X; 1057–58), without any effort to consult the regency.

Stephen, who had succeeded to the papacy while abbot of Monte Cassino, summoned Peter Damian from the monastery of Fonte Avellana to become cardinal-bishop of Ostia. The election of a pope whose brother had been a rebel against the regency suggests that the strong ties that had bound the reform movement to the empire had been somewhat weakened. At the same time, the position of the reformers in Rome was also weakened. When Stephen died in 1058, the Roman nobles supported the election of Bishop John of Velletri as Benedict X (1058–59), thereby attempting a return to the pro-aristocratic and pro-Roman policies of Benedict VIII and Benedict IX. The reformers were forced by these circumstances to seek support from the empress Agnes. Their candidate, the Burgundian bishop of Florence, Gerard, was installed on the papal throne as Pope Nicholas II (1059–61).

The reign of Nicholas has long been recognized as pivotal. Not only did Nicholas preside over a major shift in the papacy's relations with the Normans but he also issued a decree in 1059 regulating future papal elections that began the process of concentrating electoral powers in the hands of the cardinals. The foundation for a change in papal relations with the Normans had already been laid in the necessities that arose after the papal defeat at Civitate in 1053. Leo IX had earlier appointed Humbert of Silva Candida as archbishop in Sicily and now entrusted the conquest of Calabria and Arab-dominated Sicily to the Normans with the provision that they should be held from the papacy as a hereditary fief. This assertion of papal overlordship, often read solely in political terms, represented an effort on the part of the papacy to ensure its claims to jurisdiction over the churches of the still unconquered lands; it should be considered in light of the claims by Constantinople. Rather than viewing Nicholas as the author of a dramatic diplomatic revolution, it is better to place his initiative in the context of the changes that confronted the papacy from Civitate onward. The Norman alliance safeguarded papal interests in the south, ensured a measure of stability in Rome during a period of imperial impotence, and promised that independence which the reformers had worked for in their notion of *libertas ecclesiae* (*i.e.,* immunity from secular control and jurisdiction). But the weakness of the empire also led the papacy to seek support in northern Italy.

The pivotal reign of Pope Nicholas II

The Investiture Controversy. The kingdom of Italy, a creation of the Lombards, had all but ceased to exist as a separate entity in the early 11th century. Pavia no longer functioned as an administrative centre after 1024, when the royal palace was destroyed. The great beneficiary of the new situation was Milan, whose archbishop, Ariberto (Heribert) of Intimiano (Antimiano), played the role of kingmaker in the selection of Conrad II as Italian king in 1026. But the archbishop faced considerable opposition from his lesser vassals, the *secundi milites* (vavasors), who revolted on his return to Milan after supporting Conrad in Burgundy. The roots of this revolt lay in a dispute between two ranks of Milan's warrior elite, the *capitanei* and the *secundi milites,* over the inheritance of fiefs. Conrad was able to restore peace between these factions in 1037 by the Constitutio de feudis, which made the fiefs of the *secundi milites* hereditary. The settlement, however, did not create a lasting peace. A group of *secundi milites* and lower clergy led by Arialdo and Erlembaldo directed their opposition against the archbishop, who was supported by the *capitanei.* These dissidents, known as *pataria,* most probably meaning "rag pickers," also had ties to the reform movement. There were also disorders in a number of other cities. Nearby Brescia, for example, forced its bishop to flee the city. In Florence, John Gualbert, one of the leaders of the monastic reform movement, opposed the city's bishop, an admitted simoniac (*i.e.,* a person guilty of using money to obtain clerical office). Yet the unrests were too varied to fit a simple explanation. The experience of Lucca, for example, differed substantially from that of Florence. The successor of Nicholas II in 1061, chosen by the cardinals, was Bishop Anselm of Lucca, Pope Alexander II (1061–73), one of the strongest leaders of the reform movement among the Italian bishops. His position at Lucca was due in large measure to the support of Countess Matilda of Canossa, a principal figure in Italian politics throughout the period. She provided the necessary weight to support papal reform policies in northern Italy and enabled the papacy to remain somewhat aloof from a risky alliance with the Normans.

The threat posed to the imperial order in northern Italy by the cities and the crucial role of the reform papacy and its ally, Countess Matilda, evoked a strong response from the imperial party at the death of Pope Nicholas. The Lombard bishops and their allies called upon the youthful Henry to provide a successor. Even though the reformers had already selected Anselm of Lucca as Alexander II in accordance with the election decree of 1059, Henry proceeded to appoint Cadalo, bishop of Parma, who took the name Honorius II. The Cadalan schism brought together segments of the Roman nobility and the Lombard bishops, who were opposed to reform. The empire, from being a partner in reform, was emerging as the enemy of reform. Under Alexander II, Hildebrand, former secretary of Pope Gregory VI and now archdeacon of the Roman church, succeeded Humbert of Silva Candida as the leading architect of reform. The papacy increasingly reached out beyond Italy in its effort to extend the influence of the reform movement. Alexander II lent his support, most probably with the advice of Hildebrand, to the conquest of England by William, duke of Normandy, in 1066. Likewise, the papacy became more fully involved in the efforts of the kings of Spain to reconquer lands from the Muslims. The internationalization of the Roman Curia, begun by Leo IX, continued to attract important leaders of the reform to Rome. Under Alexander II the papacy succeeded in moving out from the shadows of Roman domination and imperial control and onto the European stage. Against this background, King Henry reached his majority and began to assert his rights in both Germany and Italy. Alexander and Hildebrand continued papal support for the *pataria* in Milan; as noted, the reformers demanded an end to marriage of the clergy, to the buying of church offices, and to lay control of appointments to clerical benefices. The issue of lay investiture (the practice whereby secular rulers formally presented to clerics the symbols of their various offices) gained paramount importance in Alexander II's reform program. This was a major challenge to the Ottonian system, because it threatened the emperor's long-standing rights to appoint and control church officials. It drove a deep wedge between the advocates of reform and the supporters of the empire, even dividing the reformers themselves.

The Cadalan schism

From the imperial point of view, it was impossible to separate the issue of lay investiture from the changes occurring in the political life of the northern Italian cities. Indeed, the struggle over lay investiture lent legitimacy to movements directed against imperial bishops. Milan was one among a number of northern Italian cities in which a reform movement struggled to secure greater independence from imperial control; similar movements developed in the south of Italy as well. The Normans, as noted, had first been called upon to support a Lombard-led rebellion against Byzantine rule in Bari. Somewhat later, the religious confraternity dedicated to St. Nicholas, whose body had been brought to Bari from Anatolia, served to rally local sentiment. The 11th-century reform movement grew against the background of the emergence of communes, the political organizations of the towns, in central and northern Italy.

The climax of the Investiture Controversy was reached under Alexander II's successor, Hildebrand, who took the name Gregory VII (1073–85). With Gregory, the reform movement achieved its most revolutionary form, though Gregory hardly thought of himself and his contemporary reformers in such terms. Their ideal was a restoration of the church to its primitive freedom, summed up in the Donation of Constantine, a document forged in the 8th century but generally regarded as genuine throughout the Middle Ages. According to this document, the emperor Constantine in the 4th century had granted to Pope Sylvester I and his successors spiritual supremacy and temporal dominion over Rome and the entire Western Empire. Gregory and the reformers, for whom the freedom of the church consisted of a situation in which the church was free of imperial intervention and the pope able to act without restraint for the good of Christendom, used the Donation of Constantine to bolster their program of reform. That the Donation cast a long shadow over the program of Gregory VII is especially evident in the *Dictatus Papae* ("Dictates of the Pope"), a list of brief statements inserted in Gregory's register asserting papal claims. For example, the eighth title states that the pope alone can use the imperial insignia (the symbols of temporal power). Fruit of an assiduous combing of various sources, the *Dictatus* (which dates to 1075), seems to anticipate the controversies of the coming years. Certainly, it suggests the direction in which the thought of the Roman Curia was moving. The notion of a revival of a golden age also found its expression in the artistic and architectural works of the period, as, for example, in the mosaics of the basilica constructed at Monte Cassino by Abbot Desiderius and in the newly built cathedral of Salerno.

Gregory VII and the Donation of Constantine

The conflict between Gregory VII and Henry IV (1056–1106) over lay investiture was a culmination of developments in Italy that had their origins in the last years of the pontificate of Leo IX. At the Roman synod of 1075, Gregory signaled his determination to bring an end to the practice of lay investiture. There could be no doubt that this policy would have its most drastic impact on Germany and northern Italy, where the remains of the Ottonian system constituted important vestiges of imperial control. Henry IV and his counselors realized these implications and replied at the synod of Worms in 1076. Henry employed a frontal attack on Gregory, challenging the legitimacy of his election. Gregory's response was equally provocative: he excommunicated Henry, which released his subjects from their allegiance. This calculated political response aimed at undermining Henry's position with the German aristocracy. Faced with rebellion, Henry made his celebrated trek across the Alps to meet the pope before he could come to Germany. Pope and emperor met at Matilda of Canossa's castle of Canossa. There, amid the snows of winter, Henry stood for three days as a penitent until the pope received and absolved him. Henry's action at Canossa saved him temporarily, but he remained in jeopardy. When the conflict was renewed in 1080, Gregory again excommunicated Henry, who proceeded to gather his supporters. A synod at Brixen elected Guibert of Ravenna as pope under the name Clement III (1080; 1084–1100). Henry led his army into Italy and laid siege to Rome. Gregory turned to Robert Guiscard and the Normans for assistance, who drove Clement and Henry from the city but also put it to sack (1084). Gregory went southward with Guiscard and his Normans, where he died in Salerno in 1085.

Gregory's defeat did nothing to strengthen the position of the empire in northern Italy, while it drove the papacy closer to the Normans. The election of Abbot Desiderius of Monte Cassino as Pope Victor III (1086–87) illustrates this change, since Desiderius had long functioned as an intermediary between the papacy and the Normans. The election of Urban II (1088–99), formerly a monk of Cluny in Burgundy and a strong supporter of Gregory's policy, showed the continued strength of the Roman Curia's resolve while at the same time opening the way to closer ties to the Capetian kings of France as a counterweight to the empire and an alternative to the Normans. Urban was also effective in gaining support for reform among the cities of northern Italy. Yet his most dramatic endeavour was his summons of the First Crusade at Clermont in 1095, presaged by his earlier meeting with Byzantine envoys in Italy. Urban's commitment to the crusade proceeded from his desire to heal the schism between the Eastern and Western churches, to extend the papal reform program to the Eastern churches, and to forge a new alliance within Christendom against Islām, its foremost external enemy. The crusade offered new opportunities for the maritime cities of northern Italy, which for some time had been opposing Muslim power in Sardinia, Corsica, and Sicily. Urban also worked closely with Roger I, count of Sicily, to reestablish the Latin church on the island, but he came into conflict with him over the degree of direct papal control to be exercised there. The apostolic legation he granted to Roger and his son limited direct papal intervention in the ecclesiastical affairs of the island and thereby joined the reconstruction of the church to the interests of the Norman monarchy (see below). At his death in 1099, Urban had greatly enhanced the prestige of the papacy, yet the conflict with the empire remained unresolved.

Summoning of the First Crusade

Settlement of the Investiture Controversy. The settlement of the Investiture struggle finally reached under Popes Paschal II (1099–1118) and Calixtus II (1119–24) has generally been labeled a compromise. This judgment is certainly fair, but it does not do justice to the impact that the settlement had on the church as well as on civil society. The agreement reached between Paschal II and King Henry I of England, which limited the role of the king in the appointment of bishops, marked the direction for the eventual solution reached by Calixtus II and Henry V (1106–25) in the Concordat of Worms in 1122. Thenceforward the emperor was denied the right to invest prelates with the spiritual symbols of their offices; however, as their temporal overlord he retained certain rights (the ones for Germany differing substantially from those for Burgundy and Italy). An earlier attempt at settlement had proposed a near total severance of the ties between the bishops and the monarchy, and thus an end to the Ottonian system. While such an arrangement would have satisfied every aim of the reformers, it would have plunged the empire into chaos. The bishops, well aware of their role as linchpins, objected strongly to such a radical solution, and the plan foundered. But the compromise that was reached at Worms was also fraught with danger for the empire. The reformers gained much more than the actual agreement granted. The chief beneficiary was the papacy, which succeeded in freeing itself from imperial restraints. In the temporal sphere, however, and largely by accident, northern and central Italy also found itself in a new situation largely as a result of this settlement.

The Concordat of Worms

The rise of communes. The 12th century witnessed the development of communes throughout central and northern Italy. After early beginnings in cities such as Pisa and Genoa, virtually every episcopal city in the north formed a communal government prior to 1140. The origins and developments of communes are complex, and attempts to explain them in simple terms are doomed to failure. The emphasis of 19th-century liberal historians on communal revolts against ecclesiastical repression as well as the later Marxist focus on class conflict in the development of the communes have proved too narrow. They have failed to recognize both the many different causes of violence in the communes and the diversities within the communal movement. Violence was rooted in such diverse factors as the conflicting interests of ecclesiastical institutions, the complex ties of loyalty that bound men to one another and to the institutions of their society, and shifts in the distribution of power.

The Investiture Controversy focused the efforts of the higher clergy on consolidating their rights against infringement not merely by the laity but also by ecclesiastical competitors. Examples of conflict from various places and periods show how relationships changed dramatically, often within a short period. At Brescia, for example, the bishops, who contested the abbots of Leno for control of the church of Gambara, drew support from a faction of the *milites,* the landed aristocracy, as well as from the *popolo,* composed of professional people, craftsmen, and merchants. Elsewhere, local circumstances dictated other alliances. During the period in which the cities were expanding their power into the *contado* (the region surrounding the city), elements drawn from town and countryside continually struggled for control of the commune. Alliances shifted depending on the success or failure of these efforts. At Lucca, the bishop and the commune were jointly concerned about the claims of the Abbey of Fucecchio because of its ties to neighbouring Pisa. But the efforts of bishops to establish their rights in the *contado* could also provoke conflict with the commune. Much depended on the makeup of the commune, which varied widely not only from city to city but also from period to period. The relative influence of urban men of business or rural landholders depended on the size of each group within a particular community. Where one group was small, it allied itself with others. The quest for power led to shifting and, at times, strange alliances; for example, at Brescia, in the early 13th century, one faction of local magnates drew support from heretics. Even when the commune brought together various factions in sworn associations, it still faced not merely the problem of its enemies in the city and countryside but also that of the fragile nature of the coalition on which it rested. Elaborate systems of checks on power were erected, aimed at preventing dominance by any single faction. Term limitations were imposed to force changes in the ruling councils. The consuls, so named from Roman precedent, were hedged about in similar ways.

Communal government and the *contado*

From its inception, it was clear that communal government aspired not merely to political independence but also to control of the *contado.* As a result, a complex relationship developed in which the *contado* found markets for its products, offered opportunities for investment for city dwellers, and suffered oppression from urban interests. The city, in turn, offered opportunities to the people of the countryside and helped to ensure a measure of security. There is probably no way for scholars to establish the balance of benefits and disadvantages, but one may be certain that it shifted continuously, depending on region, local conditions, and the general economic climate at a given time.

It is evident, however, that the communes of northern (and central) Italy benefited from the Investiture Controversy. The ineffectiveness of imperial power in Italy during the first half of the 12th century was critical to the development of the communes and was in large part due to the disruptions caused by the struggle over investitures and the attendant political instability in Germany. These external factors, however, do not in themselves account for the rise of the communes. For that, one must turn to internal factors, particularly the dynamics among various factions and the tensions caused by the dislocation of traditional patterns of social relationships during and following the period of the Investiture Controversy, which must be viewed against the background of rising populations and increased prosperity. The effects of the policies pursued by the reformers remained a decisive force throughout the 12th century and well into the 13th century. The partisan perspectives of contemporaries vastly distorted the nature of the conflicts engendered by the dislocation of both lay and secular institutions, showing it either as a struggle for the freedom of the church from lay power or as an effort to preserve the traditional—that is, imperial—order within society. Actually, however, given the complex network of local loyalties within both church and secular society, these conflicts fragmented loyalties of both the clergy and the laity. If the old order was weakened, it was not merely a secular order that lost but an ecclesiastical order as well. Nor was it merely the unreformed monasteries and imperial bishoprics that lost; at times, communal authority emerged stronger than any ecclesiastical power in the region. Certainly, by the early 13th century this was true in Genoa, and it was soon to be the case in Milan, Florence, Bologna, and elsewhere.

THE AGE OF THE HOHENSTAUFEN

The 12th century witnessed the creation of a new political order in Italy. It was not a tidy process, however. In the south, the foundations laid by the Normans of Capua and by the Hautevilles gained strength with the conquest of Sicily from the Muslims in the late 11th century. Following the death of Robert Guiscard, his brother, Roger I, count of Sicily, went to the mainland to consolidate his position. His son, Roger II, succeeded in establishing a Norman kingdom. Recognition for it was, however, slow in coming. Roger first obtained it from the antipope Anacletus II (1130–38), and then, under conditions that revealed the weakness of the papacy before Norman power, from Pope Innocent II (1130–43) in 1139. The papacy continued to seek support from the French monarchy in order to offset growing Norman influence. On the other hand, victory in the Investiture Controversy, even though compromised, created a situation that enabled the 12th-century papacy to assume leadership of the reform movement throughout Europe. The Lateran Councils of 1123, 1139, and 1179 marked important stages in the development of the reform papacy. From Urban II on, the central administration of the church continued to grow. By the mid-12th century, about the time that the monk Gratian was compiling his

The Norman kingdom

Decretum, the most important collection of ecclesiastical law up to that time, Rome's position as a court of appeals was growing faster than its judicial machinery could possibly accommodate. The process of definition and extension of papal control over ecclesiastical matters inevitably led to conflict with secular rulers. The determination of the papacy to protect its independence and, after the death of Matilda of Canossa in 1115, to hold onto the vast inheritance she had bequeathed to the Roman church in central Italy, as well as the ties that united the papacy to the reform bishops and to many of the laity who had supported the reform in their cities, signaled changes that had taken place since the Concordat of Worms in 1122. Finally, the keystone of the new order lay in the strength, as yet untested, of the communes. Vestiges of imperial support certainly remained both in the cities and in the countryside, but the old cause had given way before the real interests that were taking shape during this period.

The emperors of the Hohenstaufen dynasty that succeeded the Salian dynasty attempted to revive the imperial position in Italy. The first efforts of both pope and emperor in the period following the Concordat of Worms were, however, based upon the assumption that something of the old relationship remained. In part, this attitude may have been encouraged by the slight attention paid to Italian affairs by Lothair II (or III; 1125–37) and Conrad III (1138–52). Lothair's efforts against the Normans were ineffectual. Moreover, he was chiefly involved in civil war in Germany. His rival and successor, Conrad III, the first Hohenstaufen king, devoted considerable energy to the Second Crusade, which had been promoted by the Cistercian monk and monastic reformer Bernard of Clairvaux. The dominant role of Bernard of Clairvaux certainly influenced the selection of his former disciple, Eugenius III (1145–53), as pope. Forced to seek refuge in France by the internal situation in Rome, where the radical reformer Arnold of Brescia stirred both feelings of independence and a demand for more extreme reforms within the church, Eugenius cooperated with Bernard in the preaching of the Second Crusade. Although the crusade was conceived originally as an enterprise to be led by the Capetian monarch, Louis VII (1137–80), Conrad III was included when it became clear that there would be large-scale German participation. The crusade fell well short of expectations, and Conrad returned to Germany in 1149 to resume his imperial program. Eugenius, who faced Arnold and the rebellious Romans and who was heavily dependent on Roger II of Sicily during Conrad's absence on crusade, hoped that Conrad's return would provide the means to reestablish papal control in Rome, but turmoil in Germany prevented the realization of his desire. Conrad died without being able to journey to Italy to receive imperial coronation. Eugenius died the following year.

The Second Crusade

Frederick I (Barbarossa). Conrad was succeeded by his nephew, the duke of Swabia, Frederick I (1152–90), whose reign marks a major reassertion of imperial rule in Italy. Frederick saw himself not as the heir to a compromise but as a restorer of the Romano-Carolingian heritage of the German monarchy.

Papal-imperial relations. The figure of Frederick I, whom the Italians called Barbarossa ("Redbeard"), has been surrounded by legend and embroidered by myth that fit the nationalistic and liberal traditions of historiography fashionable in the 19th and early 20th centuries. But this still popular image of Frederick has undergone a major revision by scholars during the post-World War II era. The move away from nationalistic interpretations has led scholars to reevaluate the significance of major events in the imperial-papal relationship in ways that better reflect the actual historical context within which they took place. For example, the Treaty of Constance of March 23, 1153, by which both pope and emperor dedicated themselves almost to a return to the former status quo in both northern and southern Italy, demonstrated the extent to which they attempted to retain essential elements of the traditional order. But events soon showed how illusory this effort was. There was in fact little trust between the papal and imperial sides. Frederick made his descent into Italy in 1154 in order to secure his coronation as emperor. His troops were few, chiefly a band of knights under Henry III the Lion, duke of Saxony. He put Milan under the ban of the empire for refusing to answer charges laid against it by Lodi, Pavia, and Cremona. But he could do little else. He moved quickly to Rome, where a new pope, Adrian IV (1154–59), the only Englishman ever to hold the papal see, had succeeded Pope Anastasius IV (1153–54). Adrian had little choice but to continue the arrangements made at Constance, though he and his chief adviser, Cardinal Roland Bandinelli (who later succeeded Adrian as Pope Alexander III), wanted to prevent Frederick's reassertion of imperial claims to participate in papal elections. Still, they needed his support to quell the continuing unrest created by Arnold of Brescia. The emperor captured Arnold and turned him over to the prefect of the city, who hanged him, burned his body, and scattered his ashes in the Tiber River. Frederick, however, did not move against the Normans, although King Roger II of Sicily had died, and Adrian concluded a treaty with King William I (1154–66) of Sicily in 1156. Frederick's first Italian trip thus served chiefly to demonstrate the impossibility of the kind of restoration Frederick had envisaged in the Treaty of Constance, but that did not mean that he was prepared to surrender the rights of the empire. Quite the contrary, it helped to move the issues into a new arena.

Frederick's first Italian trip

Perhaps no more dramatic expression of the nature of this change could be imagined than the event that took place at Besancon, where the cardinals Bernard of San Clemente and Roland met with Frederick in October, 1157 and delivered a letter from Pope Adrian. The pope reminded Frederick of his imperial coronation and informed him that he wished to confer great *beneficia* on him. The term, which could mean either favours or, in a more specific sense, offices, was translated into German by Frederick's imperial chancellor Rainald of Dassel as fiefs, which implied that the emperor held the empire from the pope by feudal tenure. This caused an uproar among those present, particularly since Cardinal Roland went on to ask: "From whom then does he receive the empire . . .?" Although Pope Adrian denied the interpretation made by Rainald, the damage was done. More importantly, however, this incident shows that contemporaries were quite aware that they were treading on new ground. Frederick firmly rejected any implications of papal overlordship and asserted that he held the empire "from God alone by the election of the princes." That his policies were grounded in political realities is confirmed by his actions in 1158, when again he set forth to Italy. This time he sought neither a rapprochement with the papacy nor a return to the old order. He came as a ruler intent on restoring order in his domains. Having humiliated Milan, which had attempted to oppose him, he met with the cities on the plain at Roncaglia to define the royal regalia (rights) on the basis of customary law. Four Bolognese lawyers joined 28 urban representatives in this task. The text of the three laws issued at Roncaglia, however, shows the increasing influence of Roman law at Frederick's court.

Roncaglia

Institutional reforms. Already in the second half of the 11th century, studies of Roman law underwent a revival in Bologna under the influence of the jurist Irnerius and his school. Earlier emperors had, in fact, employed Roman law in their judgments and legislation. But Frederick was more conscious of its importance as a source justifying imperial actions. He issued a special privilege for scholars studying law, the so-called "Authentica Habita" (*c.* 1155), and played a leading role in the gradual evolution of the law schools in Bologna into a university. Roman law, however, was merely one source that contributed to the development of more clearly defined social and political institutions in the 12th century. The profound changes occurring in Italy in this period made innovation inevitable. Communes everywhere were experimenting with new political forms, often concealing their novelty behind traditional names. In the Norman kingdom, the effort to fashion royal institutions for disparate regions and populations led not merely to a layering of administrative institutions within the royal court but to a great diversity from one region to another. Everywhere, attempts at reconciling widely divergent legal and customary arrangements made it evident that legal uniformity was desirable; the view that each group should live by its own law no longer served the needs of society. More and more, Roman and canon law provided sources useful for reconciling differences. Frederick Barbarossa saw himself as an agent of unification. Although he represented the traditional order and was so viewed by his numerous enemies in Italy, he identified himself with the changing order that was emerging in the mid-12th century. Roncaglia was a new constitutional statement in spite of its conservative reliance on regalian right.

Northern Italy. Roncaglia laid the foundation for a new regime in northern Italy. The regalia produced an enormous income from such sources as minting coinage and collecting tolls, making Frederick the wealthiest of all European monarchs. This led him to reconceive the position of Italy within the empire. The aggressive policies pursued by him and his officials could not help but alarm Milan and its allies, and they also caused profound concerns for the papacy. The death of Adrian IV in 1159 revealed a division among the cardinals. A pro-imperial group supported Octavian of Monticello, while the opposition chose Cardinal Roland of Siena. Amid angry recriminations, the two claimants assumed the papal title, Octavian as Victor IV (1159–64) and Roland as Alexander III (1159–81). At a poorly attended council in Pavia, Frederick abandoned his neutrality and supported Victor IV. But Victor found little support beyond the boundaries of the empire. Even Bishop Eberhard of Bamberg, one of Frederick's closest advisers, made his acceptance of Victor conditional on his gaining recognition from the whole church. Nevertheless, Frederick's support was sufficient to give life to the schism. Alexander III relied chiefly on France, England, the Lombard cities allied with Milan, and the Normans in southern Italy and Sicily. On the whole, Frederick's position was the weaker one, but he refused to make concessions. Failing to win over Louis VII of France, he journeyed once again to Italy, only to face strong opposition from the patriarch of Grado, who organized the anti-imperial League of Verona. When Victor IV died in 1164, Rainald of Dassel arranged for the election of the strongly imperial Paschal III (1164–68). But Alexander III's position was also not without difficulties. The controversy between King Henry II of England and Archbishop Thomas Becket threatened to deprive Alexander of English support. Moreover, feeling increased uncertainty about his position in France (where he had fled in 1162), Alexander returned to Rome in 1165 and sought support from the Normans. The death of William I of Sicily in 1166 prevented the realization of this goal, and he was forced once again to look to France as his mainstay.

The League of Verona

Frederick sought a decisive solution in Italy in 1166. He marched with a large force to Rome, but a devastating outbreak of malaria among his troops while in the city put an end to the emperor's plans. In 1167, the Lombard cities formed a league to defend against Frederick's expedition, which included Milan, Venice, Padua, Mantua, Brescia, and Lodi.

Despite his setback, Frederick was determined to stay the course. Indeed, by this time, he could hardly turn back without accepting a near total surrender. Failing to muster support in Germany, Frederick was forced to rely on the limited resources left to him. On May 29, 1176, he met his enemies at Legnano in northern Italy. The army of the Lombard League, under the leadership of Milan, and Frederick's army engaged in a pitched battle, in which the supporters of the empire were thoroughly routed. Accepting the realities created by his defeat, Frederick waged a diplomatic campaign that secured the remarkable treaty concluded with Alexander (whom he now recognized as pope) and the Lombard communes in Venice in July 1177. This agreement settled little definitively, but Frederick obtained a six-year truce with the Lombards and was able to hold onto the Mathildine lands in Tuscany for 15 years. He restored his position in Germany and recovered from the losses endured in Rome. In 1183 Frederick converted the truce of Venice into the Peace of Constance, in which he renounced the regalia claimed at Roncaglia but preserved the administrative rights of the crown. From defeat, he thus managed to salvage a considerable portion of his imperial power.

The battle of Legnano

Frederick launched his final expedition to Italy in 1184, where he met with Pope Lucius III (1181–85). He also witnessed a diplomatic turnabout on the part of the Norman ruler, William II (1166–89), who espoused his aunt Constance, the daughter of Roger II of Sicily, to Henry VI (1190–97), the son of Frederick Barbarossa. Although Constance was not expected to inherit the throne since William and his queen might still have children, the implications of the agreement were nonetheless momentous. The papacy found itself faced with an intolerable situation. Frederick, now aiming to build a power base in Tuscany instead of Lombardy, attempted to annex the Mathildine lands. Although he failed in this, he secured the *spolia* and regalia of vacant bishoprics and abbacies from Pope Clement III (1187–91). Yet Frederick did not live to consolidate this effort. The defeat of the crusader army at Ḥaṭṭīn in the Holy Land in July 1187 and the subsequent fall of Jerusalem sent a great shock through the West. Frederick responded by taking the cross; the kings of England and France followed suit. Frederick Barbarossa was drowned in the Saleph River in Anatolia on June 10, 1190. The crusade was able to save 'Akko (Acre) and assure the continued crusader presence in the East, but it left Jerusalem in Muslim hands.

Economic and cultural developments. Frederick was the giant of the 12th-century Italian stage. He lived through a period of dramatic social and economic changes. Genoa, Pisa, and Venice became international powers during this period, with commercial interests stretching from northern Europe to Africa and the Levant. The growth of population in both town and countryside was marked by an increase in public works, ranging from town walls to canals. The growing complexity of economic organization was reflected in the development of guilds and confraternities. Even the smallest communities had their professional elite of judges and notaries alongside the nobles, merchants, and craftsmen. The vigour of the economy found its expression in the construction of new and larger cathedrals, the one at Pisa being among the most notable. Overseas trade and investment increased domestic wealth, leading to the embellishment of cities.

Culture, in turn, produced its own coin. In the Norman south, medical studies developed in Salerno. Although the kingdom of Sicily did not become a major centre for the transmission of Byzantine and Islāmic cultures to Europe, as did Spain, it nonetheless played a significant subsidiary role. Ash-Sharīf al-Idrīsī, the famous Arab geographer, dedicated his three major works on geography to Roger II of Sicily. George of Antioch and, later, Eugenius the Admiral, were important translators of Greek works into Latin. Capua, Monte Cassino, Benevento, and Salerno contributed to the Latin cultural tradition from their own rich patrimonies. Historical writing flourished in the hands of Amatus of Monte Cassino, Romuald of Salerno, Geoffrey Malaterra, and Falco of Benevento. Already in the 11th century an international clerical culture had emerged in the writings of reformers such as Humbert of Silva Candida and Peter Damian, and it grew under the influence of figures such as Bernard of Clairvaux and John of Salisbury. On the local level, Roman civic culture found its expression in clerical circles around the great basilicas and in secular circles around the prefect and, later, the senators. The north produced an early harvest of the civic spirit in the annals of the Genoese Caffaro and his successors. Although imperial themes often found a place in these cultural developments, underlying loyalties were local. Only slowly did signs of an international lay culture—largely under French influence—emerge. By the late 12th century, the whole of Italy had undergone a major economic and cultural transformation that was to provide a rich basis for the 13th century.

Roman civic culture

Henry VI. The death of Frederick's eldest son, Frederick of Swabia, on the crusade brought to the German throne his second son, Henry VI (1190–97), who had stayed behind in Germany. Thus, strangely, the son who had not expected to become king, who was also husband to a princess who had not expected to inherit a throne, found himself in a position to claim both the German and the Sicilian crowns. In Germany, the strength of Henry's support and the prestige of his father made succession certain, the more so because he was able to defeat his father's enemy, Henry the Lion, and hold his sons as hostage. But the Sicilian heritage of Constance was another matter. The nobility of the kingdom supported the popular Tancred of Lecce (1190–94), as did the English king, Richard I the Lion-Heart, the old ally of King William II of Sicily. But Henry had been able to secure a promise of imperial coronation from Pope Clement III prior to his death, and his successor, Pope Celestine III (1191–98), who deliberately delayed matters by engaging in negotiations with Henry, nonetheless proceeded with the coronation on the day following his own consecration. Henry immediately turned his attention to his wife's Sicilian inheritance. But an outbreak of typhus forced him to abandon his plans and return north. Constance herself was captured and held in Salerno. Pope Celestine declared for Tancred and recognized him as king. In Germany, much of the Rhineland joined Richard the Lion-Heart and Celestine against Henry. But the capture of Richard on his return from the crusade strengthened the emperor's hand; Henry demanded an enormous ransom and conspired with King Philip II of France to keep Richard a prisoner. When Henry finally reached an agreement with Henry the Lion in the spring of 1194, the way was open for his return to Italy.

Henry's Italian policy has been subject to considerable speculation among scholars. Some have suggested that he embarked on a program of world domination. But such a view is too grandiose. Henry was essentially a practical man. He was also an opportunist. Lacking sources that would provide real insight into his thinking, one can only conclude that he was aware of the policies of his father and that his own aims were extensions, with some modification due to changed circumstance, of those policies. The crown of Sicily was the chief new element. In 1194 he returned to Italy and subjected the Sicilian kingdom to his rule; Tancred had died shortly before his campaign began. Aided by the Pisans and Genoese, Henry entered Palermo and was crowned as king on Christmas Day. Constance had remained at Jesi, where she gave birth to a son named Constantine, to be known as Frederick Roger (later Frederick II) in honour of his paternal and maternal grandfathers. Henry devoted himself to establishing German control over the bureaucracy of the Sicilian kingdom. He planned to integrate its administration into that of the empire, employing imperial *ministeriales* for this purpose. These were originally servants of unfree origin, who had risen to become important administrators in the imperial government of the Hohenstaufen. Henry gave the trusted ministerial Markward of Anweiler the duchy of Ravenna and the march of Ancona as hereditary fiefs, thereby ensuring that the land route between the kingdom of Italy and the kingdom of Sicily was in safe hands. These measures reveal the centralizing goals that were at the heart of his vision. He tried to ensure that the German (*i.e.*, imperial) crown would be hereditary in his family, a plan that was on its way to realization when, amid his preparations for a crusade, he was felled by typhus in Messina and died on Sept. 28, 1197. In Germany, the Hohenstaufen future rested with the efforts of Henry's younger brother, Philip of Swabia, to secure the succession for Frederick Roger. In the kingdom of Sicily, Constance succeeded immediately and moved to assert her authority.

Otto IV. In northern Italy, Henry had endeavoured to preserve the gains that Frederick Barbarossa had made. But Frederick's departure on crusade and Henry's own concern with the kingdom of Sicily permitted the communes to recover from the reassertion of imperial control after the Peace of Constance in 1183. Henry's death and the ensuing imperial election bought still more time for the communes. The empress Constance withdrew her son as a candidate for the German throne and entrusted him to the regency of the newly elected pope, Innocent III (1198–1216), before she died in 1198. The German election was contested by Philip of Swabia and Otto of Brunswick, son of Henry the Lion of the Welf house. But when on June 9, 1208, Philip was assassinated by Otto of Wittelsbach in a private quarrel, Otto IV (1208–1215) emerged triumphant and began preparations for his coronation, which took place on Oct. 4, 1209. Despite his concessions to the pope, Otto had no intention of dropping imperial claims in Italy; the practical aims that had driven the Hohenstaufen impressed themselves on his policy as well. When the new emperor made clear his intention to move on the kingdom of Sicily, Innocent had no choice but to excommunicate him (1210). With the advice of King Philip II of France, the pope transferred his support to the young Frederick, thus paving the way for his accession to the German throne.

Excommunication of Otto IV

Frederick II. *Relations to the papacy.* The youthful king of Sicily knew little peace during the years following the death of his mother. Though Innocent III was nominally his guardian, Markward of Anweiler attempted to control the child-king, basing his claim to the regency on Henry VI's last will. After Markwald's death in 1202, Frederick was caught between factions in the kingdom. Only after his marriage to Constance of Aragon in 1209 did his position improve. Then, in 1211, Innocent III turned to him as his candidate for the German throne. This dramatic reversal on the part of the pope and his seeming willingness to jeopardize what most historians have viewed as the papal position in Italy has raised serious questions. True, Innocent exacted a promise from Frederick to maintain the German kingdom separate from the kingdom of Sicily, with his son Henry to be king of Sicily while he became emperor. But acceptance of such a solution raised the question whether the papacy was really committed to any long-term policy, at least one that was consistently dedicated to the separation of the Sicilian kingdom from the German empire. In any case, Frederick spent the next eight years in Germany pursuing and consolidating his position as head of the German kingdom. The kingdom of Sicily was again thrown into turmoil by the competing factions among the nobility, by the efforts of towns and cities to assert greater independence, and by growing tensions between Sicilian Muslims and their Latin neighbours. In northern Italy, imperial authority atrophied; in most places, it was little more than a hollow shell. Even in Germany, the new king found that he needed to gain the support of the German church through a broad grant of privileges. In addition, Frederick took the cross at Aachen in 1215, aligning himself with the plan of Innocent III for a new crusade.

The positive tone in the relationship between Frederick and the papacy during these early years has often been missed by historians concerned chiefly with the papal-imperial conflicts of the period. The keynote of these years was papal-imperial cooperation, especially directed toward the successful completion of a crusade. The limited results of the Third Crusade and the bitter fruit of the Fourth Crusade, which had led to the capture of Constantinople and large parts of the Byzantine Empire, had prompted Pope Innocent III to undertake a major revision of papal involvement in the crusades, culminating in the decree *Ad liberandam* ("To free the Holy Land") at the Fourth Lateran Council in 1215. Innocent's program required a level of commitment never before achieved, especially in financial and logistical terms. Frederick's assumption of the cross demonstrated his support for Innocent's program but raised problems, chiefly because of the uncertain state of affairs in Germany even after the defeat of Otto IV by Philip Augustus at Bouvines in 1214 and outstanding issues in northern Italy. With the death of Innocent III in 1216, many thought the papal plan for a crusade was in jeopardy, but his successor, Honorius III (1216–27), worked energetically to maintain Innocent's schedule as much as possible. Honorius pressed Frederick and other leaders of the crusade to hasten their preparations. Although the first contingents from the north were prepared to leave in 1217, Frederick was not among them, partly because of his failure to estimate correctly his capacity to resolve his problems in Germany and Italy and partly because of the speed of preparations for the crusade. It was not until 1220 that Frederick came to Italy for his imperial coronation and to reenter his Sicilian kingdom.

Ad liberandam

His repeated postponements of his departure for the East, granted by a cooperative Honorius, ultimately prevented him from joining the crusade. The defeat of the crusaders in Egypt in 1221 delivered a serious blow to the policy of cooperation initiated by Innocent III and nurtured by Honorius.

Nevertheless, it was a notable period of cooperation. Even among the various Italian cities, mediation became an increasingly important tool in settling differences. While Frederick made few concessions, Honorius showed that he was willing to trust in Frederick's promises to reach a final settlement of the Mathildine lands and to keep the administration of the kingdom of Sicily separate from that of the empire. Frederick promulgated imperial laws against heresy, based on the decrees of the Fourth Lateran Council. Following his coronation, he began the task of restoring royal authority in the kingdom of Sicily. His Assizes of Capua (1220) set forth a program designed to regain control of royal rights alienated since the reign of Henry VI. He also began to put in place a more effective central administration. He worked to secure the support of important members of the ecclesiastical hierarchy, including the abbots of Monte Cassino and La Cava as well as the bishops of the kingdom. He built support among the great nobles, especially the counts of Aquino, but he also encountered considerable resistance from a large segment of the nobility. Faced with this opposition and a revolt among the Muslims of Sicily, he was again compelled to postpone his participation in a new crusade, though he was careful to send aid and troops. The death of Honorius III in 1227 and Frederick's cancellation of his departure for the East because of illness broke the dam of recriminations and distrust that had been building. The new pope, Gregory IX (1227–41), excommunicated Frederick, citing as reasons his repeated postponements and his abuse of the rights of Sicilian churches during vacancies.

Excommunication of Frederick II

The kingdom of Jerusalem. An excommunicated Frederick embarked for the East, where he negotiated an agreement with the sultan Al-Malik al-Kāmil of Egypt for the return of Jerusalem on terms somewhat less favourable than the sultan had earlier offered the crusaders in return for Damietta. Frederick, who had married the heiress to the kingdom of Jerusalem in 1225 and who had an infant son Conrad from this marriage, laid claim to the kingdom. He set up a regency and embarked on a program to strengthen royal administration. In the meantime, Gregory IX, claiming provocation by the imperial vicar Reginald (or Rainald) of Spoleto, gathered an army and invaded the kingdom of Sicily. Frederick returned from the East, defeated the papal forces, and reached an agreement with the pope at Ceprano in 1230 that did much to restore the basis for cooperation. He was at last free to devote his efforts to Italy.

The Sicilian kingdom. The kingdom of Sicily was Frederick's first priority. It had long suffered neglect from his absence and internal strife. The Constitutions of Melfi, or *Liber Augustalis,* promulgated by Frederick in 1231, was a model of the new legislation developing from the study of Roman and canon law. Although it was long thought that Gregory IX opposed this body of laws, that view has been recently shown to be incorrect. In fact, he wrote to congratulate Frederick on his work. The intent of this legislation was to bring together the disparate elements within the kingdom and to knit them more effectively into a unity under royal leadership. Provisions were made for improvements in royal administration, greater efficiency in the courts, and a rationalization of civil and criminal procedures in the interests of justice. Frederick also worked to promote the general welfare of his kingdom. In 1224 he founded the University of Naples. His legislation then dealt with medical education and licensing, public health, and air and water pollution. But he did not lose sight of the place that the kingdom occupied within imperial thinking. Increasingly in the 1230s he was drawn into affairs in northern Italy and Germany that made him conscious of the importance of the Sicilian kingdom as a base for his imperial power. Very possibly, circumstance played a greater role than ideology in forcing this conclusion upon him.

The Constitutions of Melfi

The war in northern Italy. As a part of the settlement reached between him and Honorius III at the time of his coronation in 1220, Frederick had arranged to have his son Henry crowned as king of the Romans (*i.e.,* ruler of Germany) while retaining the imperial, Italian, and Sicilian crowns for himself. Henry, encouraged by some people at his court, embarked on a policy that threatened Frederick's relations with the German aristocracy. As a result, Frederick moved against Henry and placated the German nobles with his Constitution in Favour of the Princes (1232). Yet, whereas Frederick was willing to trade away some of his authority in Germany, he was determined to assert imperial rights in northern Italy. The Lombard cities, as early as 1226, had renewed the Lombard League. While Frederick was dealing with the problems caused by the Henry's rebellion between 1233 and 1235, the Lombards grew increasingly restless. Frederick gave reality to their fears by his decision to summon a diet for Piacenza in 1236 to impose his imperial authority on them. His action thus made clear his lack of interest in papal efforts to arrange compromises between him and the Lombards. Moreover, the emergence of his chancellor, Pietro della Vigna, as his chief spokesman signaled a shift away from the quiet diplomacy between emperor and papacy that he had carried on with the aid of Hermann von Salza, the grand master of the Teutonic Order, a man respected by the pope and the Roman Curia. Pietro's rhetoric was well fashioned for a propaganda war. On his side, Gregory appointed the strongly anti-imperial Cardinal James of Palestrina as his new legate in northern Italy and blocked Frederick's planned diet. In his propaganda, Frederick portrayed himself as the champion of orthodoxy working to prevent the spread of heresy in Lombardy, thus building on the theme of the cooperation between him and Honorius III aimed at stopping the growth of heresy. Gregory's rhetoric appealed to papal claims based on the Donation of Constantine and expressed his earlier concerns about Frederick's abuse of ecclesiastical rights.

Renewal of the Lombard League

The growing rift between Frederick and the papacy was not merely a revival of the papal-imperial conflict of the 12th century, though it certainly had elements in common. It had its immediate roots in the failure of the policy of cooperation employed under Innocent III, Honorius, and even Gregory himself. There is every indication that Frederick valued this relationship, but he came increasingly to see it as an obstacle to securing his imperial rights in northern Italy. On its side, the papacy had also worked to preserve good relations. But fear of Frederick's policies in northern Italy evoked memories of Frederick Barbarossa in the minds of some members of the Curia. Above all, neither the emperor nor the pope could turn back the clock on the status of the communes. In fact, from the very outset, Frederick seemed more a pawn of the emerging forces in northern Italy than a restorer of the ideal of empire. The new forces were represented above all by two tyrants, Ezzelino and his brother, Alberigo, from the ancient da Romano family, who were working to expand their lordship from their base in Verona at the expense of towns such as Padua and Brescia. Frederick relied on them for support, and in so doing he provoked the opposition of earlier supporters, such as Marquess Azzo d'Este, who now sided with the Lombards. These were the potential beneficiaries of Frederick's military activities, more so than the emperor himself.

Buoyed by early success in northern Italy, Frederick returned to Germany. He even hoped to repair his differences with Gregory, who proved amenable. However, the attempted settlement broke down. On Nov. 27, 1237, Frederick, back in Italy, dealt the Lombards a heavy blow in the battle of Cortenuova. He followed his military success with a strong 9ropaganda attack, chiefly directed against Gregory IX. But the victory won at Cortenuova proved difficult to convert into permanent gains. Milan continued to hold out. In the following summer Frederick laid siege to Brescia but failed to take the city. Gregory excommunicated the emperor, citing causes in keeping with previous papal criticisms. What was at stake, however, was not some ideological high ground but recognition that Frederick had violated the rights of the church in the

The battle of Cortenuova

kingdom of Sicily. Gregory attempted to use the machinery developed for the crusade to the East to gather money and manpower to oppose Frederick, who, in turn, warned his fellow rulers of the danger posed by these efforts. The papacy accused Frederick of failing to support the crusade mounted by Thibaut of Champagne in 1239 and delaying its departure for the East. Gregory wished to recall him to the program on which the papacy had been insisting since the reign of Innocent III, but Frederick's own concerns were with his European domains. It was not that he opposed the papacy's desire for a crusade; he wanted to settle matters in Italy first.

The factors shaping political factions. The breach between emperor and pope that marked the remainder of the reign of Gregory IX and that grew more intense under Innocent IV (1243–54) was undoubtedly a major factor in shaping political factions in northern Italy throughout the 13th century. But the notion that it was the conflict over church and state that determined political developments exaggerates its role. As already noted, local and regional factors underlay the politics of the northern communes. The conflict of religious and political ideology emerged chiefly in the second half of the 13th century, but later debates often hearkened to the vituperative papal-imperial propaganda of the 1240s. In some respects the lines had already been drawn, at least in part, by the internal political disputes that had begun to dominate Italian urban life from the mid-11th century on. The role played by the reform movement in the emergence of political factions is known only in part, but its importance cannot be denied. However, it would be a mistake to view this influence solely in terms of a clerical-lay dichotomy. The emergence of papalist and imperial parties that later in the century called themselves Guelf and Ghibelline, respectively—based on terms taken from the divisions between the Welf house of Otto IV and the Hohenstaufen (Waiblingen) house of Philip of Swabia and Frederick II—echoes the struggle over rights. The term *pars ecclesiae* (party of the church), which became more common in the second half of the 13th century, has generally been viewed as a reference to support for the papacy, but it also referred to support for local churches. Both meanings of the term were correct, and the earliest usages seem to favour a more local interpretation.

Communal developments

The changing character and composition of communes often followed the fortunes of this struggle over rights. Increasingly, divisions between magnates and *popolo* were a mask for genuine political realities that concealed the process of coalition-making characteristic of early 13th-century urban politics. The regime of the *podestà* (which had its origins in imperial appointees), formed in the second half of the 12th century to provide greater stability and protection against violence, was already becoming more professionalized in the age of Frederick II. The emphasis on professional officials, strongly evidenced by mid-century in the writings of Albertanus of Brescia and others, marked a reaction to the military aggrandizement of an Ezzelino da Romano or Azzo d'Este and a defense of communal values.

The *pars ecclesiae* was very often the party that controlled the commune and stood for communal independence. Although some disputes with bishops were an inevitable feature of the Italian urban scene, the alliance between bishops and communes grew more common in the 13th century. Imperial ideology was largely driven from the field. Likewise, disputes of a more economic nature, based on class, varied in importance from one place and time to another; but, on the whole, such factors were subsidiary to the deep-rooted concerns over rights, especially over property, on both the local and imperial-papal level, which served as driving forces in the shaping of Italian urban politics.

The end of Hohenstaufen rule. The final decade of Frederick II's reign marked the end of the imperial system in Italy. Although Frederick seemed at times on the verge of repeating Barbarossa's achievement, he was unable to marshal the resources needed for the task. His kingdom of Sicily fell more and more victim to his need for money to fight his war in the north, which all but ended the efforts at good government that had motivated his Constitutions of Melfi (1231). Increasingly Frederick mortgaged his income to Roman bankers and subordinated the interests of the kingdom to the needs of the empire. His relations with the northern Italian maritime powers, save for Pisa and to some degree Venice, continued to deteriorate in the 1240s. But he remained dependent on income from taxes on their trade and could not entirely break ties with them. Instead, he tried to compete directly with them and to control their access to the ports of the kingdom. The pressures of war largely determined his policies. The death of Pope Gregory IX in 1241 at first promised the possibility of a resolution. But it soon became clear that the cardinals were unable to proceed to an election because of their internal division. It was only in 1243 that they finally selected the Genoese cardinal, Sinibaldo Fieschi, one of the leading legal experts of the period, who took the name Innocent IV (1243–54). The new pope soon demonstrated his unwillingness to compromise with Frederick, identifying papal interests closely with those of the Lombards. Under threat from the Romans, he left Italy and summoned a general council of the church to meet in Lyon in 1245. This council sealed the rupture between Frederick and the papacy and ensured that there would be little chance for a reconciliation in future negotiations. Frederick's position remained militarily secure, but threats of conspiracies and growing opposition marked the final years of his reign. He was the last medieval emperor to assert his rule in northern Italy.

Innocent IV

With the death of Frederick II at Fiorentino in Apulia on Dec. 13, 1250, the stage was set for major changes that were to affect the future of Italy for centuries to come. The papacy determined that it would no longer tolerate the rule of the Hohenstaufen in Italy and opposed Frederick's son and successor, Conrad IV, as well as Frederick's natural son, Manfred, who became de facto ruler in the kingdom of Sicily and, following Conrad's death in 1254, secured the crown for himself. Conrad's son, Conrad V (Conradin), continued, however, to be the official heir. Even before Innocent IV died in 1254, the papacy tried to secure aid from the English king Henry III (1216–72), promising the Sicilian crown to his younger son Edmund. Innocent's successor, Alexander IV (1254–61), soon discovered the folly of this effort and began to search elsewhere. In the meantime, Manfred consolidated his position in the kingdom and in the march of Ancona. In northern Italy the removal of Frederick opened the way for tyrants such as Ezzelino da Romano to assert their virtual independence under the guise of supporting Manfred. Even the death of Ezzelino in 1259 did not impede the trend toward military rule. Almost everywhere, the ethos of strong government overwhelmed the rule of the communes.

Charles of Anjou

The death of Frederick and the virtual failure of his line left a power vacuum. The papacy turned for support to Charles of Anjou, brother of King Louis IX (1226–70) of France. While Frederick was alive, Louis had remained aloof from the conflict and had even expressed his disapproval of the actions taken by the papacy against Frederick. However, after the death of Conrad, he supported Charles of Anjou's bid for the Sicilian kingdom against Manfred. At the same time, Charles' arrival in Italy set a new direction for politics in the north, where he was welcomed by the Guelfs. He proceeded to Rome, followed by a crusader army that was supported by the pope, Clement IV (1265–68), himself a Frenchman. Charles fended off Manfred while awaiting his own army, which had traveled overland. On Feb. 26, 1266, he defeated Manfred and his forces on the plain of Grandella near Benevento. This battle ended Hohenstaufen rule in Italy and began the Angevin dominance that lasted through most of the rest of the 13th century. Charles succeeded in Naples and asserted his leadership in Rome and in northern Italy. He quickly defeated a futile effort by Conradin and had him executed in Naples (1268).

The papal party of the Guelfs was dominant in the north; the papacy ruled in the centre; the Angevins were in control in Sicily and Naples; and the whole was held together by the strong personality and ambitious dreams of Charles of Anjou. As matters turned out, many of his dreams differed but slightly from their imperial counterparts. Moreover, despite their aggressive and grandiose character, most of them seem to have had little basis in reality. Charles had designs on the Byzantine Empire and North Africa, as is evident from Louis IX's futile crusade to Tunis in 1270. The north of Italy was growing more and more restive under the rule of Charles. Pope Nicholas III (1277–80), sensing the direction of things, removed Charles from his position as senator at Rome and from his vicarship in Tuscany. The papacy also spearheaded efforts at reconciliation between Guelfs and Ghibellines. But Charles resisted and endeavoured to increase his influence over the papacy. Pope Martin IV (1281–85), a strong Angevin supporter, restored him to his former offices. But before Charles could reap the benefits from his reinstatement, he was faced with an insurrection in Sicily. The so-called Sicilian Vespers, an uprising on Easter Monday of 1282, when citizens of Palermo attacked the French garrison, led to a protracted war known as the War of the Sicilian Vespers. The king of Aragon, Peter III, came to the aid of the rebellious Sicilians, while Charles was indirectly aided by his nephew, the king of France, as well as by the papacy. After 20 years of intermittent warfare, the outcome was the division of the kingdom founded by the Normans. The mainland of southern Italy continued to be held by Charles' successors (the Angevin dynasty), whereas the island of Sicily came under the hegemony of an Aragonese dynasty.

The Sicilian Vespers

As the 13th century came to an end, Italy had not yet recovered from the disasters of the age of Frederick II. Milan was dominant among the Lombard communes and was soon to make its bid for hegemony in northern Italy. Florence, however, had increased in size, wealth, and power to become the dominant force in Tuscany and the chief bulwark against the ambitions of the Visconti, a powerful Milanese family. The papacy vacillated under Angevin influence, while the kings of Naples carried on a vain war to retake the island of Sicily from the Aragonese. External influences, especially from the direction of France, increasingly dominated the political life of Italy, culminating, at the turn of the century, in the disastrous conflict between Pope Boniface VIII (1294–1303) and King Philip IV the Fair of France (1285–1314). Italy had escaped imperial dominance only to fall under the shadow of the French. Perhaps there is a certain irony in the country's exchange of one heir of Charlemagne for another.

Italy in the late 12th century.

Economic developments. The 13th century witnessed an enormous increase in prosperity not only in the maritime cities but also in the growing centres of the cloth industry, especially the woolen textile industry in Tuscany. Venice came to dominate the rich trade with the Byzantine Empire, especially after the Fourth Crusade (1204). Genoa, which eclipsed Pisa in the latter part of the century, expanded its trade in the western Mediterranean and Provence. During the second half of the 13th century, Florentine influence, benefiting from close ties with the Angevins and the papacy, prospered in the kingdom of Sicily. Italian cities came to reflect the new wealth. By the end of the century, the first mansions of the rich, although small by later standards, began to adorn the cities alongside new municipal buildings and the churches of the mendicant orders, especially of the Franciscans and Dominicans.

Cultural developments. The culture of this period reflects both the new wealth and the tensions it generated in Italian society. The legal culture, which had flourished in Bologna in the 12th century, came of age in the 13th century and spread to other parts of Italy. Lawyers formed a large educated class with an appreciation of rhetoric and grammar and some taste for poetry, history, and philosophy. In the first half of the century the court of Frederick II was an important centre for these studies, as is evident in the letters of Pietro della Vigna, the emperor's chief spokesman. The chronicle of Riccardo of San Germano proved the best that the century would produce. Frederick's court also attracted such figures as Michael Scot, whose translation of mathematical and scientific treatises from Arabic into Latin made Sicily an important centre for their transmission. Frederick's own study *De arte venandi cum avibus* (*On the Art of Hunting with Birds*) drew not only on earlier writings but also on his own and his contemporaries' observations and experience. The incipient Dominican *studium* in Naples produced Thomas Aquinas, arguably the greatest thinker of the age. Frederick, however, did not continue the rich Norman tradition of mosaic art and architecture, best represented by the Cappella Palatina in Palermo and the cathedrals of Cefalù and Monreale. Instead, Frederick was more noted for his castles, especially his starkly beautiful Castel del Monte in Apulia.

The legal culture

Thirteenth-century Rome saw a flourishing of the arts in sculpture and in the stone work of the Cosmati and in the mosaics that adorned the walls of such churches as San Paolo fuori le Mura and Santa Maria in Trastevere. By the end of the century, Arnolfo di Cambio, whose work in Florence was to gain him greater fame, produced important sculpture in Rome. But Rome was chiefly the centre of the papacy and of an international clerical culture. Although the papal chancery grew apace in this period, producing thousands of polished letters in its distinctive style, other studies found little place there. It was more devoted to practice than to study.

Secular legal studies, grammar, and rhetoric took deep root in the north. The influences that were streaming across the Alps from France and which were later to draw numerous Italians, including the Florentine scholar Brunetto Latini, northward influenced the writings of such figures as Albertanus of Brescia. In Milan, Bonvesin da la Riva, poet and eulogist of his city, composed his *De magnalibus urbis Mediolani* ("Concerning the Great Works of the City of Milan") in 1288. At Padua, Rolandino reacted against the incursions of Ezzelino da Romano in his *Chronicle*. While in exile from Florence, in the early 1300s, Dante Alighieri, the greatest of all Italian poets, completed his towering epic poem, *La Divina commedia* (*The Divine Comedy*). Dante's literary art found its visual equivalent in the brilliant frescoes of Giotto di Bondone in Padua (Arena Chapel), Florence, Assisi (in the Magdalen Chapel in the lower church of San Francesco), and Naples (destroyed). (J.M.Po.)

Italy in the 14th and 15th centuries

CHARACTERISTICS OF THE PERIOD

The failure of the Hohenstaufen emperor Frederick II and his successor kings of Sicily to dominate Italy in the course of the 13th century left the peninsula divided among a large number of effectively independent political units. The impotence of efforts by rulers from beyond the Alps to impose their authority upon it was clearly and finally demonstrated by the expedition (1310–13) of Henry of Luxembourg, crowned as Emperor Henry VII. An idealist who believed that as God's secular vicar he had a divine mission to restore peace to "the garden of the Empire," Henry entered Italy in 1310 with the consent of Pope Clement V (1305–14) and seemed at first to prosper. He sought, as an honest broker, to reconcile Guelf (*i.e.*, pro-papal) and Ghibelline (*i.e.*, pro-imperial) factions, but it was soon apparent that any attempt to override those old loyalties implied a massive assault upon the political status quo, a revolution that would be fiercely resisted. Florence, in particular, saw as unacceptable not simply any concession to its enemies but any restoration of imperial power.

Henry VII's Italian expedition

In these circumstances Henry was increasingly driven into exclusive alliance with those who were opposed to the Guelfs and became himself merely a leader of a faction. As a result, both the papacy and King Robert of Naples, who had originally favoured his coming to the peninsula, returned to their traditional anti-imperial stance. The dream of peace by imperial fiat dissolved, and Henry turned to war. But his death from fever at Buonconvento, near Siena, in August 1313 was to break the hopes of the imperialists forever. Later emperors who intervened from the north—Louis IV the Bavarian (1327–30) and Charles IV of Bohemia (1354–55, 1368–69)—came with much more limited aims, not as universal monarchs but as short-time players on the Italian scene, seeking there such limited gains as, for instance, the prestige of imperial coronation at Rome. However much these emperors preserved their formal de jure claims to rule, any imperial central authority in Italy had disappeared. In its place stood a complex, often chaotic grouping of many rival powers whose hostilities and alliances fill in wearisome detail the pages of contemporary chroniclers.

This reality has to be seen together with other divisions in a peninsula that was characterized by sharp regional differences in climate, land formation, economic development, customs, and language. (A 13th-century chronicler praises a contemporary as a skilled linguist because of his fluency in "French, Lombard, and Tuscan." There was to be no common literary language before Dante, and then only in verse, not prose.) These very pronounced diversities have led many commentators to dismiss as futile an attempt to construct a general unified history of Italy in this period and to insist that the only coherent synthesis that can be formulated is one based upon its constituent parts. For these authors the only true history will consist of separate accounts of the six major powers—Sicily, Naples, the Papal State, Florence, Milan, and Venice—together with those of some 15 to 20 minor powers—such as Mantua, Montferrat, Lucca, and Siena—which were interspersed between them. (This ignores the ambiguous case of Genoa, economically extremely powerful but politically pitifully weak.)

Campanilismo

There is much in such contentions. It would be unwise to play down the overwhelming spirit during the 14th and 15th centuries of *campanilismo* (local patriotism; the spirit of "our campanile's taller than yours"). Only a minority of people living at that time could ever have heard the world "Italia," and loyalties were predominantly provincial. It is true that among certain classes, such as merchants who traveled beyond the Alps or scholars who looked back nostalgically to Roman republican or imperial glories, some elements of national consciousness survived. Dante, seeking in his *De vulgari eloquentia* (written 1304–07; "On the Eloquence of the Vernacular") to find, amid what he described as "a thousand different dialects," "the elusive panther" of some basis for a common vernacular literary language, argued that there were some "very simple standards of manners, dress, and speech by which our actions as Italians are weighed and measured." However vague this claim may appear, one can certainly see in the peninsula some elements that, taken together, made a strong contrast to the world beyond the Alps: a common legal culture, high levels of lay education and urban literacy, a close relationship between town and country, and a nobility who frequently engaged in trade.

Yet ultimately one must conclude that any interest or importance that is to be attached to this period springs above all not from any "national" considerations or reflections upon the Italian peninsula as a unity but rather from three particular features that were witnessed in, at least, some parts of it. These were, first, the maturing, often in the face of severe opposition, of that remarkable economic development which had originated in earlier centuries. Though shaken in the course of the 14th century, northern and central Italian trade, manufacture, and financial capitalism, together with increasing urbanization, were to continue with extraordinary vigour and to have remarkable influence throughout much of the Mediterranean world and Europe as a whole—a development that served as the necessary preliminary for the expansion of Europe beyond its ancient bounds at the end of the 15th century. Second, in parallel with this, came the extension of de facto independent city-states, which, whether as republics or as powers ruled by one person or family (*signorie;* singular: *signoria;* ruled by *signori,* or lords), created a powerful impression upon contemporaries and posterity. Finally, and allied to both these movements, it was from this society that was born the civilization of the "Italian Renaissance," a Renaissance that in the 15th and 16th centuries was to be exported to the rest of Europe.

ITALY TO C. 1380

The southern kingdoms and the Papal State. Not all regions were to witness favourable economic or constitutional development or to receive anything but reflected rays from the sun of the Renaissance. In the south the Sicilian Vespers of 1282 had separated for more than 150 years the island of Sicily from the kingdom of Sicily (which until then had consisted of both the island and the southern mainland). On the mainland thenceforth, the successors of King Charles of Anjou (d. 1285) ruled as vassals of the papacy. Normally described by contemporaries as "Kings of Naples" (though resolutely continuing to call themselves "Kings of Sicily"), they pursued a 90-year war against the kings of (island) Sicily. That war, ultimately unsuccessful, was financed by harsh taxation of the only productive element in the kingdom—namely, its impoverished workers-on-the-land. This extension of royal fiscalism, already oppressive at the time of the Norman kings, fixed the region in wretched poverty and destroyed all possibility of native capitalist growth. As a result, during the 14th century almost all trade and banking came into the hands of northern Italians, particularly Florentines. At the same time, outside a few restricted areas (Sulmona, coastal Puglia, Campania) that produced considerable surpluses of grain, an arid climate and inferior soil made for poor agricultural development.

Against this background, feudal disorder flourished. Under King Robert I (reigned 1309–43; known to his literary flatterers as "Robert the Wise"), who made no less than five attempts to conquer the island of Sicily, the monarchy was able to resist the more extravagant demands of the nobility for rewards for their military and political support. But, with the accession of Robert's granddaughter Joan I (1343–82), royal authority withered away, court factions dominated, and civil war (1347–52) followed. Quelled at this point, baronial turbulence revived at the end of Joan's reign in a conflict between two branches of the Angevin family (those of Durazzo versus those of Provence) which claimed recognition as heirs of the queen. The eventual victor, King Ladislas (1386–1414), benefiting from the turbulence provoked by the Western Schism (see below), was able to boast of considerable military success in central Italy and was even able to gain what to some observers seemed to be a brief position of predominance in the peninsula. But the accession of his sister, Joan II (1414–35), inexperienced and childless (without, that is to say, obvious heirs), brought a renewal of anarchy to the Neapolitan kingdom, in which true power was held not by the monarchy but by a few powerful owners of vast estates (latifundia) who were allied to the monarchy through blood or service. Below these barons existed a large number of petty nobles with minuscule fiefs; and still lower was a mass of workers-on-the-land, living close to subsistence.

Trinacria

Meanwhile, the island kingdom of Sicily—or Trinacria, as it was often called—was ruled from 1296 to 1409 by a cadet branch of the royal house of Aragon. This house, in rebellion against its feudal overlord, the papacy, and engaged in constant war with its northern neighbour, went through a pattern of monarchical weakness and economic decline similar to that shown by the Angevins of Naples. With the death of King Frederick III (1337), the island, which hitherto had been weakly feudalized, was now increasingly divided up by substantial concessions of royal lands to a grasping baronial class. Of particular importance in this group were the three great families of the Ventimiglia, the Chiaramonte, and the Passaneto—men so powerful that contemporaries described them as "semi-kings," having below them some 200 lesser, poor, and violent feudatories. In these years, with an economy dominated largely by Catalan merchants, Trinacria looked to Aragon and its great port of Barcelona rather than to the peninsula to the north.

If the southern kingdoms limped through the 14th century in internal strife and economic backwardness, so, too, did the Papal State lying to the north of the Kingdom of Naples. In March 1303 Pope Boniface VIII, in conflict with King Philip IV of France over papal jurisdiction, had been seized at the papal residence of Anagni by a small band of French and Roman adventurers. Though released almost immediately, he died a month later of, it was said, deep humiliation. The Papal State had been founded in order to preserve the independence and spiritual authority of the papacy, yet here, clearly, it seemed to have failed. Partly because of the menacing Roman baronage, partly again through the pressure of the French king, Pope Clement V decided to abandon the peninsula and seek refuge at Avignon. Here between 1307 and 1377 the papacy was to reside in greater safety. Italy was now "bereft," as Dante, who witnessed these developments, testified, "of its two suns," of both the papacy and the empire.

The effects of that withdrawal were twofold. First, the "lands of the Crucified One," as the church dramatically described its territorial state, were reinforced in their secular anarchy, and everywhere local "tyrants" seized power from papal officials. Yet, at the same time, the traditions of the church inevitably required that the papacy should return to that Rome where St. Peter had, it was said, preached and suffered. Hence, over the years, with alternating enthusiasm and lack thereof, the French popes struggled sporadically to establish obedience, peace, and control over their Italian lands. These efforts, indeed, set the agenda for a great part of the foreign politics of the Italian states in the period. Notable were the attempts at reconquest of the Papal State by Cardinal Bertrand du Poujet (1319–34) and Cardinal Gil Albornoz (1353–63). Yet the results were slight. After heroic expenditure of money and blood, Albornoz was to attain some measure of order, largely by appointing the more amenable tyrants as "papal vicars" and by securing from them promises of payment of taxes and services in return for acknowledgment of overlordship. But even these muted successes were to be unstable. With the outbreak of war between the papacy and Florence in 1375, most of the vicars cast off their allegiance. Three years later the Papal State was cast into even greater disarray with the outbreak of the Great Western Schism (1378–1417). For almost four decades, until the Council of Constance, unity was shattered by rivalries between popes and antipopes—one French, one Italian, and later a third one.

Amid the confused struggles that engulfed the Papal State in this period, one incident in particular stood out for men of the day and excited the imagination of posterity. The city of Rome, deserted by the papacy, presented a sombre picture of shepherds, herdsmen, labourers, and artisans dwelling by ruins that testified to past glory and were now taken over as the residences of powerful aristocratic families. The Colonna, Orsini, and Annibaldi established their fortifications amid the remains of the Mausoleum of Augustus, the Forum, and the Colosseum, and from there they fought out their ancient rivalries. Here in the 1340s rose the remarkable figure of Cola di Rienzo. A notary, the son of an innkeeper, possessing an imagination that easily accepted the most flattering fantasies, he gained esteem from the rumours he circulated that he was the son of Emperor Henry VII. An avid reader of classical history and an interpreter of ancient inscriptions, intoxicated by the past splendours of Rome, he preached to his fellow citizens the recovery of its former greatness. Inspired by the Lex Regia, the supposed right of the Roman people to confer authority on the emperor, he announced that the citizens of his own day, under his leadership, could assume that right and resolve all disputes between rival claimants to the office. Achieving prominence as the most eloquent member of an embassy dispatched to Avignon to complain of the absence of the papacy, he excited the admiration of many (including the poet Petrarch) at the papal court. On his return in May 1347, with the help of some mercenary soldiers, he seized power in the city and, by a parliament summoned at his command, was awarded the title of "Nicolai, the Severe and Clement, the Tribune of Freedom, Peace and Justice, and Liberator of the Holy Roman Republic."

Cola di Rienzo

The following month Cola invited all the Italian states to appear before him to discuss "the security and peace of Italy." It is a remarkable testimony not so much to his eloquence as to their desperate wish for peace that no less than 25 communes answered his call. During a remarkable round of ceremonies, in the presence of the communes' representatives, Cola announced that the Romans held jurisdiction over the whole world and conferred Roman citizenship upon all citizens of other Italian states. These chimerical pretensions (described by a contemporary as "fantastic stuff which won't last long") very soon came to be unveiled as such. In the following December, faced with an increasingly suspicious pope and a Roman citizenry satiated by novelties, Cola was driven from the city. He returned to Rome and was appointed senator in 1354 (essentially a puppet of Albornoz' attempt to dominate the Papal State), but within less than three months he faced a popular revolt that ended with his death. Cola's importance lay not so much in anything he had achieved as in the demonstration of how powerful an influence the thought of classical Rome could exercise on men of the time. He survives in cultural history (as hero, for example, of the German composer Richard Wagner's opera *Rienzi*) and in the myths (certainly no more than myths) that he had planned the unification of Italy and was a prophet of the 19th-century Risorgimento.

The *popolo* and the formation of the *signorie*. Meanwhile, in the course of a long process extending through the 13th and 14th centuries, within the towns of the Papal State and most towns of northern and central Italy, there arose from the old communes a new form of government, that of the *signoria*. The communes of the 13th century had become increasingly dominated by the conflicts of the nobility entrusted with their governments. These divisions, though often moved by the Guelf and Ghibelline parties, were, in fact, largely driven by personal, economic, or quite local political rivalries—all inflamed by ideals of chivalric honour and an everyday acceptance of the traditions of vendetta. In large part as a response to these conflicts, there had arisen within the communes the movement of the *popolo*—*i.e.*, of associations of nonnobles attempting to win a variety of concessions from the nobility.

Membership of the *popolo*

Within the ranks of the *popolo* were, in the first place, those who, made wealthy by trade, banking, exercise of a profession, or landholding, sought to gain entrance to the ruling noble oligarchies. The second group was composed of prosperous members of the artisan or shopkeeping classes who, while not normally seeking a direct position in government, were attempting to obtain a more satisfactory administration of the finances of the commune (particularly a more equitable distribution of taxation), a greater voice in matters that most directly concerned their interest (for example, the licensing of the export of food), and, in particular, the impartial administration of justice between noble and nonnoble. Above all, the *popolo* (like many of the nobility themselves) had an interest in the creation of a civic order that would end violent party conflicts and lessen the effects of noble vendettas.

In some towns the *popolo* movement succeeded in bringing about constitutional change. In those communes where the nobility did not monopolize all wealth and where the development of trade, industry, and finance had brought into being a complex social structure, the existing oligarchies were persuaded to come to terms. This came about more easily when the *popolo* succeeded in ending party struggles so violent that they could be described as a form of civil war. Here, often against the background of some disaster, such as defeat in war, it became normal for a council of the *popolo*, under a captain of the *popolo*, to be established side by side with the old council of the commune under its podesta as a consultative element in what was now described as the government of "the commune and *popolo*." In Florence, where the movement enjoyed its greatest success, the *popolo*, organized in seven major and five lesser guilds, was established in 1282 not simply as the partner of the commune in government but as the dominant element within it. Moreover, in January 1293, by the Ordinances of Justice, it declared that the members of 152 powerful families were to be deemed "magnates" and, as such, excluded from personal participation in government and subject to particular disadvantages in law vis-à-vis nonmagnates.

Nevertheless, in all but a few towns, the *popolo* proved unable to solve the problem of public order, and in these circumstances "the peaceful and tranquil state" of the cities came instead to be established by *signori*, who were powerful party leaders. From the second half of the 13th century, having triumphed over, destroyed, or permanently exiled their opponents, these men began to give institutional form to their power and to pass it on to their sons as a hereditary right. What they offered in return to their subject citizens was the hope of eliminating anarchic civil violence by the exercise of superior force. It was in this way that, in the course of the 14th century, *signoria*, or permanent legal rule by single families, began. From the communes the *signori* would obtain their titles, the authority to control the communes "according to their own will," and the right to pass on this grant to their chosen successors. With the passage of time these usurped legislative trappings lent the appearance of legitimacy to their rule. By the end of the 14th century the *signori* normally sought some legitimization of their power by obtaining authorization from the emperor or pope to act as "vicars" over the territories their families had come to rule. As such, during the 15th century these hereditary lordships—or, in effect, principalities—seemed to constitute the natural order in large areas of northern Italy.

Signorie in northern and central Italy

So, in the Trevisan March, Verona fell to the della Scala (or Scaligeri) family in the 1260s, as did Vicenza from 1311, while Padua was subject to the Carrara (or Carraresi) from 1328. In Lombardy the Bonacolsi and then, from 1328, the Gonzaga family came to be sole rulers of Mantua, while the Visconti achieved *signoria* at Milan from 1311. During the next 35 years the Visconti extended their lordship and built a very large territorial power over Cremona (continually from 1334), Pavia, Lodi, Bergamo (1332), Como (1335), Piacenza (continually from 1337), Tortona, and Parma (1346). In Emilia the Este (Estense) family, already established at Ferrara from 1264, extended their power to Modena (1288) and Reggio (1290). In the northern sector of the Papal State the towns of Romagna and the Marche fell to *signori* between 1315 and 1342; with the failure of Cardinal Albornoz' attempts at reconquest, most of its territories came to be granted out in vicariate. Thus, in the hundred years from the middle of the 13th century, northern and central Italy had witnessed a profound transformation in constitutional forms, political life, and attitudes toward authority. The rule of a city-state by one man was no longer seen as a strange and temporary expedient but as a normal aspect of life. Under the new regimes the councils of the communes and *popolo* still remained, but their role was limited to minor administrative tasks or to formal approval of the political decisions of the *signori*. Essentially, all that was left of the old communal system was its administrative service, a core of skilled notaries who kept the machinery of government in operation. Meanwhile, in return for their absolute power, the *signori* brought an end to political anarchy, restored or created harmony within the upper classes of the towns, and reconciled the interests of the *popolo* and nobility.

Nonetheless, the emergence of the *signorie*, however important, was only one element in the constitutional history of the northern and central Italian towns in the 14th century. It was a movement largely confined to the Veneto, Lombardy, Emilia, the Marche, and, in a more feudal form, Piedmont. In most towns of Umbria and Latium the papacy was able to prevent their establishment. In Tuscany they were very largely unsuccessful. Lucca fell to *signori* in the first half of the 14th century, notably with the rule of the remarkable Castruccio Castracani between 1316 and 1328, but the town experienced a strong revival of republican government from 1369 to 1392. Florence underwent only brief interludes of signorial governance. Some neighbours of Florence—Volterra, Prato, Pistoia, San Gimignano—were conquered by it before any *signoria* could be established in them. In Liguria, Genoa was continually unstable through the violent conflicts of its noble houses. Rather than submitting itself to any one family, the town oscillated between communal government and a series of *popolo* life-dictatorships (of which the most memorable was that of Simone Boccanegra, future hero of an opera by the Italian composer Giuseppe Verdi.) Two communes, Siena (at least in the 14th century) and Venice, rejected signorial government entirely.

It would, therefore, be misleading to think of Italy in the 14th century as being dominated exclusively by *signori*. As well as the new principates, there were some communal governments—including those of Venice and Florence, two of the most powerful cities in the peninsula, which both survived and developed into powerful territorial states with very strong republican traditions. In part this may be ascribed to the fact that it was much more difficult to seize control over a patrician oligarchy of bankers and merchants than it was to dominate a society consisting of landowners, artisans, and rural workers. Societies with highly developed economies were much less amenable to princely control. In republics an economy that would be menaced by internal disunity, and a ruling class united at least in its pursuit of commercial advantage, provided a powerful stimulus to the preservation of public order and to the repulsion of any individual or family seeking political domination.

Venice in the trecento (14th century). In was, in fact, in the 1290s and the hundred years that followed—broadly speaking, the same period in which the *signorie* were consolidating their position—that the two principal republics established and secured the essentials of those constitutions which were to last (in the case of Florence) into the 16th and (in the case of Venice) even to the late 18th century. In this period the dominion of Venice included the islands of the lagoon and the *dogado*, a thin strip of the mainland around it, together with an overseas empire consisting of most of the Dalmatian coast, the island of Corfu, various islands in the Aegean, the coasts of the Peloponnese, and Crete. The overseas territories were often valuable in themselves, and they served also as a series of staging areas for Venetian commerce. In the establishment of the city's constitution, the key moment came in February 1297, when the Great Council, in which sovereignty resided, was expanded to take in more than 1,000 members. From that time forward, and in particular from the 1320s, admission to that body became more difficult, and from the 1390s it ceased altogether. What has been traditionally described as "the Closing (Serrata) of the Great Council"—that is, the creation of an oligarchic government dominated by a fixed hereditary caste—did not occur in 1297 but actually extended over some 100 years from that date.

The Great Council

From its members the Great Council elected the Senate, the Lesser Council, the Quarantia (judiciary), the

Council of Ten, all the committees of government, and, by an extremely elaborate system of lot, the doge. The "Excellent Lord, by the Grace of God, Doge of Venice, Duke of Dalmatia and Croatia, the Most Serene Prince" enjoyed immense prestige and had an important role in the coordination of government. Yet, however powerful his personality, he was increasingly bound by his councils. In reaction against such control, the doge Marin Falier, from one of the greatest families of the city, conspired in April 1355 to overturn the sovereignty of the nobility. He paid for the attempt with his life. A century later (1457) the patriciate were to show their strength again when Francesco Foscari was deposed by the Council of Ten. As a political figure the doge came to be diminished; henceforth he could promote his own initiatives only by submitting them to the Great Council and the Senate.

The Venetian Senate was originally composed of 60 members (the Pregadi), but by 1450 it had grown to 300 members, of whom some 200 had the right to vote. It was particularly concerned with foreign policy, war, and matters of commerce. Its chosen ambassadors constituted the first and finest diplomatic service in Europe. In addition to their regular dispatches, these men were expected on their return from duty to produce a detailed account (*relazione*) of the government and country to which they had been attached. The Senate also organized the fleets and the recruitment and supervision of condottieri (see below); it controlled the markets of grain, salt, wine, and oil; and it built the principal merchant galleys and organized the regular convoys (*mudae*) in which they sailed to "Romania" (Constantinople and the Black Sea), "Flanders" (London and Brugge), and Tunis. The Senate auctioned the lease of ships for each voyage, nominated their master mariners, and laid down elaborate regulations for their crewing and equipment.

Council of Ten

Side by side with the Great Council and the Senate stood the Council of Ten. In 1310 Baiamonte Tiepolo and other nobles had sought to seize power from the dominant faction in the Great Council. It was after the suppression of this conspiracy that the Council of Ten, armed with exceptional powers, was created. At that time it was given a limited life to watch over the security of the state, but, after the attempted coup by Marin Falier in 1355, the office became permanently established. Its members controlled the secret police, espionage, and counterespionage, and they exercised in some measure their own judicial system. From that power base they came to exert a strong influence on financial and diplomatic administration.

Below the patrician class, who formed and monopolized all the political offices of the Venetian state, existed a less privileged class, that of the citizens. Consisting of about 2,500 males of the status of notaries and the like, they controlled the civil service. Their leader, the grand chancellor, though not a patrician, was, as head of the civil service, one of the most important men in the republic. Outside the ranks of the citizens were the disenfranchised majority of the population—labourers, shopkeepers, artisans, and, in great numbers, seamen.

Under this constitution, the Serenissima ("Most Serene Republic") produced the most satisfactory form of government and society known in the world at that time. Petrarch's praises of "the most miraculous city of Venice, rich in gold but richer in fame, strong in power but stronger in virtue, built on both solid marble and the harmony of its citizens, secured more by the harmony of its citizens than by its surrounding seas," echoed a virtually universal praise. Such rhetoric, typical of most discussions of the republic from the 13th to the 16th century, gained its persuasive power from the real social concord that Venetian government, as none other, did indeed provide. This outstanding success at home was matched by victories abroad. In the second (1294–99) and third (1351–55) Genoese-Venetian wars, the Genoese, the Venetians' principal economic rivals, gained numerous victories against the republic, and in the fourth war (1378–81) they were temporarily able to seize Chioggia and Malamocco on the lagoon at the heart of Venice's power. Yet, overall, with the superiority of its state structure and civic spirit, Venice was always able to win the wars. In the overseas empire a careful administration secured from its peoples, if not passionate loyalty, at least a submission that drew no small strength from the threat of an alternative Turkish dominion.

Florence in the trecento. In Florence, the other great republic of northern Italy, the key constitutional moment came in 1293 with the Ordinances of Justice. Though modified somewhat two years later, it ensured the continuance of a system in which sovereignty was acknowledged to rest with the *popolo,* an elite class drawn from the seven major guilds, or *arti maggiori*—that is, the judges and notaries, the Calimala (bankers and international traders in cloth), the money changers, the silk merchants, the doctors and apothecaries, the wool merchants, and the dealers in furs. Together with dominant figures from five guilds of lesser importance (the *arti medie,* or middle guilds, consisting of the butchers, the shoemakers, the smiths, the stonemasons, and the secondhand dealers), the *popolo* gathered every two months to elect six priors who ruled Florence as supreme magistrates.

Ordinances of Justice

Behind these forms, the men who effectively ruled were members of the *popolo grasso* ("fat people"), consisting of bankers and businessmen of great wealth, who professed allegiance to the Guelf party. Yet the survival of guild government was, in these years, often precarious. The dominant faction was often sharply divided by fierce rivalries. So in 1302 the "Black" Guelfs, in alliance with Pope Boniface VIII, succeeded in expelling the "Whites." Among the White Guelfs at this time was Dante Alighiere (1265–1321), who had taken some share in public office. Doomed to spend the rest of his life in exile, he was to write *La commedia* (*c.* 1308–21; later named *La divina commedia* [*The Divine Comedy*]), whose pages still offer eloquent testimony to the extreme bitterness of domestic conflict in these years. Moreover, external pressures forced the city to accept the lordship between 1313 and 1322 of King Robert of Naples and then again, between 1325 and 1328, of Robert's son, Charles of Calabria. It was perhaps fortunate for the continuance of the commune that Robert was too preoccupied with his own kingdom to establish any full and permanent control and that Charles died prematurely.

Yet it can be plausibly argued that, despite such political difficulties, Florence reached the apogee of its prosperity during the first three decades of the 14th century. Its population grew to about 95,000 people, and a third circle of walls, constructed between 1284 and 1333, brought it to encompass an area it was not to surpass until the middle of the 19th century. In the 1290s construction began on the new cathedral (Duomo) of Santa Maria del Fiore (the dome was not completed until 1436) and the fortress-residence of the Palazzo Vecchio—both potent symbols of the commune, to which, shortly, was added a third, Giotto's campanile.

Up to the beginning of the 1340s, Florence reigned supreme in long-distance trade and in international banking. From that time, grave shocks struck its economy, and these, combined with failure in war, led to another brief experiment in signorial rule; in 1342 a protégé of King Robert, Walter of Brienne, titular duke of Athens, was appointed *signore* for one year. Almost immediately on his accession, Walter changed this grant to that of a life dictatorship with absolute powers. But his attempt to ally himself with the men of the lower guilds and unenfranchised proleteriat, combined with the introduction of a luxuriant cult of personality, soon brought disillusion. An uprising in the following year restored, though in a rather more broadly based form than hitherto, the rule of the *popolo grasso.*

Guild rule was to continue virtually unchallenged thenceforth until 1378. In that year the overthrow of the regime came about not through the imposition of a *signore* but from factions within the ruling class, which in their turn provoked the remarkable proletarian Revolt of the Ciompi. In the wool-cloth industry, which dominated the manufacturing economy of Florence, the *lanaioli* (investor-managers) worked on the "putting-out" system; they employed large numbers of people (9,000, by some calculations) who worked in their own homes with tools

The Revolt of the Ciompi

supplied by the *lanaioli* and received wages by the piece. Largely unskilled and semiskilled, these men and women had no rights within the guild and in fact were subjected to harsh controls by the guild. In the Arte della Lana (the wool-cloth guild), a "foreign official," was responsible for administering discipline and had the right to beat and even torture or behead workers found guilty of acts of sabotage and theft. The employees, who were often in debt (frequently to their employers), subsisted precariously from day to day, at the mercy of the trade cycle and the varying price of bread. With them, among the ranks of the *popolo minuto* ("little people") or *ciompi* ("muckers" or "mates" at this low rank of society), were day labourers in the building trades as well as porters, gardeners, and poor and dependent shopkeepers. On occasion these poor, in Florence as all over Italy, were provoked to riot when bread was scarce, but they were normally powerless to organize efficiently against guilds and governments—both of which were poised to act with extreme penalties against anyone who defied their authority.

In effect, the poor rose to revolt only at the prompting of members of the ruling class. So it was in the Revolt of the Ciompi of 1378. In June of that year Salvestro de' Medici, in an attempt to preserve his own power in government, stirred up the lower orders to attack the houses of his enemies among the patriciate. That action, coming at a time when large numbers of ex-soldiers were employed in the cloth industry, provoked an acute political consciousness among the poor. In their clamour for change, the workers were joined by small masters resentful of their exclusion from the wool guild, by skilled artisans, and by petty shopkeepers. Expectation of change and discontent fed upon each other. In the third week of July, new outbreaks of violence, probably fomented by Salvestro, brought spectacular change: the appointment of a ruling committee (*balìa*) composed of a few patricians, a predominating number of small masters, and 32 representatives of the *ciompi.* Michele di Lando, foreman in a cloth factory, was appointed to the *balìa* as "standard-bearer of justice."

The *balìa*

In their six-week period of rule, the men of the *balìa* sought to meet the demands of the insurgents. Guilds were formed to give standing to their members, more equitable taxation was established between rich and poor, and a moratorium was declared on debt. Yet, angry at the slow pace of change, the poor remained restive. On August 27 a vast crowd assembled and proceeded to the election of the "Eight Saints of God's People." Then they marched on the Palazzo Vecchio with a petition that the Eight Saints should have the right to veto or approve all legislation. But by now all the temporary allies of the poor were alienated from the spirit of revolt. The rich resisted, won over Michele di Lando with a bribe, called out the guild militias, and drove the protestors from the scene.

Normality was reestablished within a few days. The new guilds were abolished, and the poor returned to that impotence which was, throughout Italy, their lot. Malnutrition quenched rebellion, leadership was lacking, and the limited horizons of their lives made any ideal of betterment short-lived. The main effect of the revolt was to introduce at the top of society a regime that was narrower and more oligarchic than that which had survived for the previous 30 years.

Economic change. Meanwhile, changes in the character of the economy in town and country profoundly affected the development of both the republics and the *signorie.* Although, in contrast to the rest of Europe, northern and central Italy in this period are often described today as driven by an "urban economy," most Italians still lived on the land, and the prosperity of any town was closely dependent upon its *contado* (the area in the countryside over which a commune ruled). Here, despite differences in farming produced by different climates and types of soil, certain patterns of development can be discerned within the peninsula. By the end of the 13th century, tenurial serfdom had virtually died away, and other forms of landholding were evolving to take its place. Sometimes peasants worked the land as freeholders (as, in fact, many peasants had always done, even at the very height of the manorial system). Sometimes (and this was particularly true of large ecclesiastical estates in northern Italy) lands were let out on perpetual hereditary lease for low rents—a procedure that, in effect, was often to lead to the virtual dispossession of church proprietors in favour of secular tenants. But the most common new tenancy from the 13th century was that in which short-term leases were offered in return for heavy rents either in money or, more often, in kind. Among such leases the one that came to figure most prominently, especially in well-cultivated land in central and northern Italy, was sharecropping, particularly *mezzadria.* In contracts of *mezzadria,* the landlord provided half the seed sown and in return received half of the tenant's fruits. Frequently the contract was renewable every year—a provision holding considerable insecurity for the lessee, who was obligated, at will, to leave the land at term. Often, in order to make sure that the landlord received a full return from his lease, detailed conditions were attached on rotation of crops, plowing, digging, and harrowing. In all, this form of tenure, which was to emerge as a central feature of northern Italian rural life up to the mid-20th century, can be seen less as an agreement to let land than one to hire labour.

Mezzadria

At the same time, a system of more intensive farming was developing. Before the mid-13th century, large homogeneous estates were a rarity, and it was very unusual for one proprietor to own half or more of a parish. From that time on, however, scattered portions were increasingly consolidated into united farms such as the *poderi* of modern Tuscany. Profits from commerce were used for building drainage, planting trees, erecting homesteads, and acquiring livestock, manure, and agricultural instruments. In these areas, the common-field system began to disappear, common pasture declined, and a growing number of individual properties were hedged in. Labourers came to live on the farm, leaving the village to house a reserve of casual workers. Thus, by the end of the 14th century, the old landscape of dispersed strips of land and fortified villages had frequently given way to broad estates dominated by country houses, leisure seats of urban landowners. Yet these developments were in no way uniform, even in the Emilian plain and in Tuscany, where they were most common. In the South the latifundia, the large estates owned by a few landowners, continued in existence, but they were now farmed with hired labour.

Evidence of *bonifiche* (drainage works) and the clearing of wasteland suggests a continual expansion in agricultural production up to the 1340s. So, too, did that period see the prosperity of trade, manufacture, and banking. Within the peninsula, communes were compelled to engage in large-scale marketing of food simply to provision the cities. For a town such as Florence, which at the beginning of the 14th century could gain from its own territories just enough food to feed its population for five months of the year, this commerce was literally a matter of survival. But, at the same time, trade in food and other bulk goods was matched by long-distance commerce in luxuries. With the decline of Pisa in the 13th century, Venice and Genoa remained the principal centres of this traffic. Venetians and Genoese had their own quarters and consulates in Syria and Palestine and at Constantinople and Alexandria. Sailing from their ports or traveling inland to Damascus and Aleppo (heads of the Asian caravan routes), they held a virtual monopoly of East-West trade, exchanging wood, steel, and arms for "spices" (the generic name for all precious goods from the East, including pepper, ginger, cinnamon, nutmeg, silk, dyes such as cochineal and indigo, cotton, drugs, and sugar).

The East-West trade

To the northeast, traders from Italy penetrated the Black Sea to draw grain, fish, salt, and slaves from the Crimea. Farther east still, they traveled to the Caspian Sea, Azerbaijan, and China. "The road from Tana [on the Sea of Azov] to Cathay," a Tuscan merchant's handbook told its readers, "is quite safe by day and by night according to the merchants who report having followed it." Indeed, in the first half of the 14th century there were Italian merchant colonies (with husbands taking their wives along) and seven Italian missionary bishops in China.

In the West, too, Italian commerce expanded its empire. The introduction of the compass to the Mediterranean, leading to new marine charts (the *portolani*, the earliest survivors of which date back to the 1290s), gave new wings to maritime daring. From at least 1277 the Genoese, and from 1314 the Venetians, carried out annual galley convoys through the Strait of Gibraltar and thence, on a compass setting, across the Bay of Biscay to the English and Flemish ports on the English Channel. Out into the Atlantic, Genoese navigators made contact with the Canary Islands. At the same time, the 14th century saw changes in shipping. Trireme galleys expanded from 50 to, sometimes, 150 tons, and Italian ports began to employ the "cog," a square-sailed vessel capable of carrying 150 slaves and ideal for bulk cargoes. Meanwhile, in banking, the most prominent of financial houses to extend their operations beyond the Alps were the Bonsignori company of Siena and the Florentine houses of the Acciaiuoli (with 53 branches throughout Europe), the Peruzzi (83 branches), and the Bardi (even larger than the Peruzzi). Within Tuscany again, beginning in the 14th century, the manufacture of textiles became a major industry.

Growth was accompanied by changes in the character of commerce. Resident capitalists appeared holding a tight grip on far-flung factors and commission agents through a network of correspondence, and with them appeared new mercantile techniques in which Italians could boast primacy—account books with Arabic numerals, double-entry bookkeeping, marine insurance, bills of lading, bills of exchange, and a mature law of the sea and law of commerce. The expansion of commerce was uneven; it was not found in the South, and even in central and northern Italy it still left most towns as no more than small market centres for the surrounding countryside. Nevertheless, the major commercial centres in the opening decades of the 14th century witnessed an economy that was never again to expand so fast and be, in relation to the rest of Europe, so powerful.

New mercantile techniques

Famine, war, and plague (1340–80). Italy's thriving economy was shortly to meet severe challenges. Among these, first, were famines, which affected most of Italy in the years 1339–40, 1346–47, 1352–53, and 1374–75, and to these catastrophes were added a general expansion and intensification of war. The 13th century saw the diffusion of the crossbow, whose bolt far surpassed the arrow of the longbow in its power to penetrate. The crossbow obliged mounted knights to adopt heavier armour for better protection. Hence arose the need for stronger and more numerous horses. Such technical developments began to make the practice of warfare much more expensive and professional, and in these circumstances mercenary troops came increasingly to supplement and then, often, to replace the old citizen militias. With the 14th century these troops were in ever larger numbers raised not through the hiring of individuals but by the drawing up of a *condotta* (contract) with a condottiere (contractor), who in time of war would engage to bring a band of up to several thousand soldiers to the aid of a commune.

Given the difficulties of securing political control over Italian military leaders (who might, it was suspected, be easily tempted to take over the state), it became common, beginning in the 1330s, to negotiate with non-Italian condottieri. Their forces rapidly grew to immense size. In the 1350s "The Great Company," founded by Werner of Urslingen, consisted of 10,000 troops and some 20,000 camp followers and had its own government, consultative council, bureaucracy, and foreign policy. Throughout the 1360s and '70s these "mobile states"—for example, the companies of the Englishman John Hawkwood and the Germans Albrecht Sterz and Hannekin Baumgarten—dominated war in Italy, and in times of peace they were all too likely to subject their former employers to a variety of blackmailing threats.

These changes in the art of war went hand in hand with a considerable expansion in the power of governments. The weak, decentralized communes of the 13th century, with comparatively primitive administration and very light taxation, gave way in the 14th century to republics and *signorie* with much stronger political control and exclusive new means of fiscal exploitation. Revenues were raised through property taxes, *gabelle* (*e.g.*, taxes on contracts, sales, transport of goods into and out of town), and forced loans (*prestanze*), while sophisticated measures, including the consolidation of state debts into a form of national debt, were devised to service long-term deficit financing. At Florence, for example, where from 1345 state debtors were issued securities of 5 percent interest, negotiable in the open market, revenues rose from around 130,000 florins in the 1320s to more than 400,000 florins in the 1360s.

Such innovations—fruit of the interrelated needs of food provision, war, and taxation—brought about considerable growth in bureaucratic institutions and in the number of administrative officials. At the same time, however, while allowing war to be waged on a larger scale, they were increasingly used to divert productive wealth into war. They were, that is, an important element in promoting those setbacks that can be observed in many sectors of the economy during the 1340s. In that decade, with trade already disrupted by the beginning of the Hundred Years' War in France, the overextension of Italian (particularly Florentine) banks became clear. In 1343 the Peruzzi company collapsed, in 1345 the Acciaiuoli, and in 1346 the Bardi.

Still more disastrous was the arrival from the East of the Black Death. Galleys and cogs brought the plague in its bubonic and pneumonic forms to Messina in early October 1347. By January 1348 it reached Genoa and Pisa, by February Venice. From these ports it spread throughout the peninsula and on to the rest of Europe. Estimates of how many died vary between one-third and one-half of the population. Yet the effects were not confined to 1348, as bubonic disease was henceforth rooted in Italy. Although slackening in its power and moving more sporadically, the disease returned to many parts of the peninsula, in both town and countryside, in 1361–62, 1363, 1371, 1373–74, 1382–83, 1398–1400, 1407, and 1410–12. Thereafter it continued as a town disease in individual, sporadic, but continually threatening assaults up to the 18th century.

ITALY FROM C. 1380 TO C. 1500

Political development, 1380–1454. From the 1380s to the 1450s Italy was distracted by a long series of large-scale wars. The principal aggressor in these conflicts was the Visconti family, who, having seized the *signoria* of Milan, had extended their power to many other cities, from Asti in Piedmont to Reggio in Emilia. From 1385 the ruthless and energetic Gian Galeazzo Visconti (created duke of Milan by Emperor Wenceslas in 1395) embarked on a series of diplomatic and military campaigns that brought him to virtual hegemony over northern and central Italy. He was assisted in this by a series of dynastic marriages—essentially, the bartering of Visconti wealth for noble blood—which gave the family immense prestige. Gian Galeazzo's first wife was Isabelle of Valois, daughter of King John II of France; his sister, Violante, was married (albeit briefly) to Lionel, son of Edward III of England; and his nieces were married to the dukes of Bavaria and Austria.

The *signoria* of the Visconti family

In 1387 Gian Galeazzo seized Verona and Vicenza from their *signore;* in 1388 he took Padua and other territories in the Veneto. These coups provoked the suspicions of Florence, and, after the failure of attempts to delineate their respective spheres of influence, three wars erupted between the two powers (1390–92, 1397–98, 1400–02). Gian Galeazzo apparently achieved an overwhelming predominance, for he was recognized as *signore* of Pisa and Siena in 1399 and of Perugia, Spoleto, and Assisi in 1400. In June 1402 he took Bologna. Florence was now encircled, and perhaps it was saved from conquest only by Gian Galeazzo's death in September from plague. At his death the state that he had built up collapsed, and his son, the vicious and incompetent Giovanni Maria Visconti (duke 1402–12), was incapable of restoring the dynasty's fortunes. With the accession of Giovanni's brother, Filippo Maria Visconti (duke 1412–47), however, a new era of Visconti expansion dawned. By 1422 Filippo Maria had restored the family's Lombard possessions. Thenceforward, until the middle of the century, there came a series of virtually continuous conflicts against an alliance of Florence and Venice.

Until the 14th century Venice had ruled only the lagoon, Venice those eastern and Adriatic possessions that had served to maintain its commerce, and, on the Italian mainland, a thin strip of land bordering the lagoon. Yet the rise of Visconti power from the 1380s persuaded the Serenissima, finally, to establish itself as a territorial power on the peninsula. If the old *signori*—the Scaligeri at Verona and the Carraresi at Padua—had seemed from time to time in the past to present a menace to the free passage of goods from Venice over the Alpine passes and into the centre of Lombardy, the threat of the Visconti dukes, with all their power, could only reinforce Venetian apprehensions. With Gian Galeazzo's death, the republic turned, accordingly, to extending its control over the mainland. Between 1403 and 1405 it took over Verona, Vicenza, and Padua. Between 1411 and 1420 the city seized the wide territories of the ecclesiastical prince, the patriarch of Aquileia in Friuli. In 1426 it conquered Brescia and in 1428 Bergamo. These acquisitions proved immensely profitable. It was calculated in 1440 that taxes from the new possessions yielded 306,000 ducats, as against 180,000 from the colonial possessions (which were, at the same time, much more expensive to defend). The "Veneto," as it came to be known, was rich, populous, and fertile—and a good market for the city's trade. In the newly subjected towns the old civic oligarchies continued to hold a measure of local power, though now under the supervision of Venetian podestas and captains. Below them, peasants and urban workers acquiesced in a system that imposed some external check upon exploitation by the town patriciates.

Venetian expansion had been effected through an alliance with its fellow republic, Florence, against Milan. Yet this entente, in part through the Venetians' very success, was shortly to disappear. On the death of Filippo Maria Visconti without male heirs (August 1447), some prominent citizens proclaimed Milan a republic. But they proved incapable of maintaining order in the state, which in 1450 surrendered to Filippo Maria's son-in-law, the powerful condottiere Francesco Sforza. Francesco was swift to proclaim himself duke. This revolution led soon to a revolution in the diplomatic alignments of the peninsula, with Florence now and for more than 40 years afterward adhering to Milan as its principal ally in its search to maintain the status quo and its own power. Following the collapse of the Revolt of the Ciompi, Florence itself had come to be ruled by a narrow oligarchic government under the personal domination of Maso degli Albizzi (from 1382 until his death in 1417) and then of his son, Rinaldo (until 1434). The Albizzi regime successfully resisted the Visconti and then a temporary threat from King Ladislas of Naples (1408–14), and it also contributed to that expansion over Tuscany which, since the mid-14th century, had transformed Florence into a "territorial state." The city had absorbed Volterra in 1361 and Arezzo in 1384; now it went on to conquer Pisa, with its port, in 1406 and to purchase Livorno from Genoa in 1421. Seeking further expansion, however, it failed to conquer Lucca in a war fought between 1429 and 1433.

That failure was largely responsible for the supersession of the oligarchy dominated by the Albizzi by an oligarchy subordinate to Cosimo de' Medici. Cosimo, who attained an unofficial personal dominance over the state in 1434, was to hold it until his death in 1464 and then pass it on to his descendants. Cosimo was the principal architect of an alliance with the Sforza of Milan that culminated in the Peace of Lodi (1454). By this pact Milan, Florence, Venice, and then (1455) King Alfonso of Naples and Pope Nicholas V bound themselves together in an "Italian League" against any power, Italian or foreign, that should disturb the existing balance of power. At the same time, the treaty established special machinery for the peaceful settlement of any disputes that might arise among the states. Despite some local conflicts, the creation of the Italian League brought about a much more peaceful era in the second half of the century. Peace was assisted, above all, by a general exhaustion among most of the major powers, whose economies and societies could no longer support the strains imposed upon them by wars.

The Italian League

The states of Italy in the quattrocento (15th century). *The southern monarchies and the Papal State.* In the south the island kingdom of Sicily under Alfonso V of Aragon (1416–58) was used mainly as a base for his ambitions upon Naples. Thereafter Sicily was governed by viceroys who subjected its interests to those of Aragon. Examples of Sicily's hispanicization are the establishment of the Inquisition (1487) and the expulsion of the Jews (1492). So, too the Kingdom of Naples, conquered by Alfonso between 1435 and 1442, pursued an unpromising development, its peace continually threatened by the rival claims of the Angevin and Aragonese dynasties. On his death in 1458, Alfonso left Naples to his illegitimate son, Ferdinand I (1458–94). Ferdinand maintained his rule only with difficulty, suppressing baronial revolt with an extreme severity that served further to alienate support.

Ferdinand, at least, was able to retain control until the days of the French invasion (1494). The Papal State, on the other hand, virtually dissolved at the time of the Western Schism (1378–1417). Southern Emilia, the Romagna, the Marche, and Umbria were given up to numerous *signori* acting as "papal vicars," among whom the most celebrated were the Este of Ferrara and the Montefeltro of Urbino. In the two cities of Bologna and Perugia, the Bentivoglio and Baglioni families, respectively, retained predominance, though without obtaining the vicariate. The church still ruled some territories directly, notably Ancona and much of southern Umbria, but in Latium its power was threatened by strong baronial families—in Rome itself antipapal and republican sentiment still survived. Not until the reign of Pope Alexander VI (1492–1503) was there to be a determined attempt to assert authority over the whole state. Until then the popes enjoyed the worst of all worlds, condemned for that deep involvement in secular politics which their position as temporal rulers had thrust upon them while, at the same time, remaining largely powerless to extract obedience from their principal vassals.

The Papal State

Venice. By contrast, Venice in the 15th century, with a population of perhaps 100,000 in the city and 1,000,000 on the mainland, enjoyed a golden age and could be considered a major European power. Its overseas empire was augmented by the inheritance of Cyprus from the French Lusignan family in 1489, and its economy was still capable of generating large profits. In 1423 the doge Tommaso Mocenigo calculated that the Venetian marine consisted of 45 state and private galleys employing 11,000 seamen, 300 large cargo vessels with 5,000 seamen, and 3,000 smaller craft employing 17,000 men. Either from the Fondaco dei Tedeschi ("Warehouse of the Germans") by the Rialto Bridge or in state-organized convoys, the precious goods of the East continued to be distributed through Europe. In industry, the state-owned Arsenal provided shipbuilding yards and dry and wet docks for the maintenance of huge numbers of vessels. Manufacture flourished, above all in silks and cottons, tanning, and, on the island of Murano, glassblowing. On the mainland, expansion continued with the acquisition of Ravenna in 1441 and of the agricultural Polesine region of Rovigo (north of Ferrara) in 1482–84. For observers throughout Europe, the "myth" of Venice excited admiration originating in surprise that so many could participate in government without having it end in anarchy. Venice's social stability continued, assisted by a legal system that strove consciously to preserve equal justice for the powerful as well as the weak and by the particular attention given, through some 120 *scuole*, or charitable organizations, to the needs of the poor.

Yet, amid general prosperity, the second half of the century witnessed three developments that foreshadowed grave future problems. First, in July 1499 Vasco da Gama returned to Lisbon from India with a small cargo of spices, threatening an end to the virtual monopolization by the Venetians of Eastern trade. Second, the Ottoman Turks, having taken Constantinople in 1453, continued their advance in Greece, the Balkans, and the Mediterranean. In the course of the first Turkish war (1463–79), Turkish cavalry raided Dalmatia and Friuli; Venice lost the strategically important island of Negroponte (Euboea, or Évvoia) and agreed to pay tribute to the sultan. At the end of the century, the second Turkish war (1499–1503) brought still greater losses. In this fight against

Italy in the 15th century and (inset) Florentine expansion.

From W. Shepherd, *Historical Atlas*, Harper & Row, Publishers (Barnes & Noble Books), New York, revision copyright © 1964 by Barnes & Noble, Inc.

the Ottomans—and here lay the third source of future weakness—the republic was hindered by the suspicions of its fellow Italian states, which were troubled by Venice's rapid expansion on the mainland and afraid that it might, in the words of Pope Pius II, "be seeking the monarchy of Italy." However untrue, it was a sentiment that struck many contemporaries, and they united in praising both the beauty of the city and the menace they detected in its power. So the French statesman Philippe de Commynes, recalling his visit to Venice in 1495, wrote admiringly of its churches, monasteries, and palaces, its 30,000 gondolas, its Grand Canal ("the fairest and best-built street, I think, in the world"), the Basilica of St. Mark, the Arsenal, and other attractions. Venice was, he thought, "the most triumphant city that I have ever seen." Yet its rulers were "so wise and so bent on enlarging their territories, that, if not prevented in time, all the neighbouring states may lament it too late."

Florence. Venice in the 15th century remained, despite all, an immensely strong power and able to preserve unimpeded the exercise of its republican constitution. In both these matters it contrasted with Florence under the Medici. The foundation of the family's fortunes were laid by Giovanni di Bicci (1360–1429), who founded the Medici bank and, in 1422, was appointed as banker to the papacy. His son Cosimo, who dominated the *reggimento* (principal patrician families) from 1434, united his vast financial resources with a keen intelligence. His natural simplicity of manner and plethora of folksy sayings were well designed to avoid offending (as far as possible) republican sympathizers. In a city proud of its traditions of "freedom," he maintained his claim to be a private citizen, refused all titles of lordship, and held the office of "standard-bearer of justice" for only three two-month periods.

The rule of the Medici family

Cosimo gained adherents by giving gifts and loans to all orders in society as well as to churches, confraternities, and religious orders and also by granting patronage to writers and artists. He was responsible for commissions to the sculptor Donatello and the architects Michelozzo (Medici Palace) and Filippo Brunelleschi (the choir and nave of San Lorenzo) and for the building of villas in the countryside at Careggi and Cafaggiolo. Founder of a great library, he subsidized the scholarship of the Neoplatonist Marsilio Ficino, the humanist Poggio Bracciolini, and the collector of antiquities Cyriacus of Ancona. In politics he moved with moderation, gradually, and sporadically. In no way could his rule be thought of as the exercise of despotic power. Cosimo was always dependent upon securing the support of a majority among the *reggimento*, who saw themselves as his allies in retaining their economic and social predominance in the state. By and large the Medici regime was acceptable to the patrician class because it stabilized those conflicts within it that had broken the unity of Florence before 1434.

Certainly Cosimo's influence was sufficient to allow his son, Piero, to take over this informal rule at his death in 1464. More remarkably, on the death of Piero in 1469, it passed to his son Lorenzo, then only 20 years old. Lorenzo's later fame as "the Magnificent" (a title given to anyone of prominence at the time) was acquired partly as a tribute to the charm of his personality, partly by a careful projection of his own image, and partly through the perceptions of Florentines of a later generation who, looking back through the dark years that followed his death, tended to think of his era as a golden age. Yet in some respects that appellation is exaggerated. In foreign politics Lorenzo made a disastrous error in the 1470s when he attempted to prevent Pope Sixtus IV from establishing a power base in the Romagna. This led to the loss of the papal account by the Medici bank and a conspiracy between members of the pope's family and the Florentine Pazzi family to overthrow Medici rule. In April 1478 the Pazzi assassinated Lorenzo's brother Giuliano but failed to kill Lorenzo, and the insurgents, denied support by the citizens, were captured and executed. Yet the "War of the Pazzi" (1478–80) that followed, with Florence pitted against a papacy allied to Naples, proved dangerous and expensive, and Lorenzo emerged from it only with great difficulty.

Following these events, Lorenzo pursued a more cautious and successful path in foreign affairs. On the death of Sixtus in 1484, he made a friend of the successor, Innocent VIII, and through this intimacy Lorenzo was able to acquire the cardinal's hat for his son Giovanni. (And it was Giovanni, as Pope Leo X, who was to ensure the triumph of the Medici throughout Tuscany in the 16th century.) But the claims made for Lorenzo as "the peacemaker" of Italy, even as a "constructor of a balance of power," have no substance—except insofar as, by virtue of ruling over a militarily weak state, he was, in his last years, willy-nilly caught in a balance of weak states from which only Venice stood out. In addition, Lorenzo (a man of genuine intellectual and aesthetic interests, who had been educated as a humanist rather than as a merchant) can be criticized as a businessman. Even with allowance for the fact that the Medici bank had to meet political as well as strictly economic ends, with loans to political allies who might be poor risks financially, it remains true that it was inadequately supervised and, for this reason, close to failure by 1492.

In the subject territories of the *contado*, Lorenzo was able to suppress any rebellion. An attempt by Arezzo to free itself from Florentine commercial exploitation in 1471 led to the sack of the town by mercenary troops in Florentine pay (though whether or not this was at Lorenzo's express will is uncertain). In Florence itself, the period of his predominance was marked among the patriciate by a more aristocratic lifestyle, expressed, among other ways, by a revival of jousting and lavish expenditure on clothes, palaces, and the arts—all of which have been seen as at variance with the older traditions of republicanism. Yet, in public patronage of the arts, Lorenzo—perhaps because he had less money, perhaps because the family's houses were already filled with works of art—did less than his father. Lorenzo's taste was directed to small, private pieces, as found in his collection of antique cameos, medals, and gems and in the pastiche-antique model statuettes produced for him by Bertoldo di Giovanni. With this went a creative interest in architecture. Lorenzo read Leon Battista Alberti's *De re aedificatoria* (promulgated 1452, published 1485; *Ten Books on Architecture*), wrote to the duke of Urbino asking to see the plans of his new palace, and entered his own design in the competition for a new facade of the Florence cathedral. Only his death (in April 1492) at the age of 43 saved the judges from what might have been a particularly difficult decision.

Milan. Within the duchy of Milan, meanwhile, the Sforza family sought to maintain its newly acquired power. Francesco (duke 1450–66) offered his subjects not only a measure of peace as well as patronage of humanism and the arts but also the disadvantages of tyrannical rule. His successor, the cruel and lustful Galeazzo Maria Sforza (1466–76), was assassinated in a conspiracy of three young men who combined personal grievances and republican sentiments. His son and heir, Gian Galeazzo, was a minor. In 1480 the regency government came under the control of Galeazzo's brother, Ludovico "il Moro," who ruled as duke from 1494 to 1499. Ludovico maintained the traditional splendour of the Milanese court and, among many other artists and engineers, employed Leonardo da Vinci (who executed for him the "Last Supper" at Santa Maria delle Grazie) and Donato Bramante (with architectural work at Sant'Ambrogio and at Santa Maria delle Grazie). Yet the prosperity of the duchy under his rule, derived from agricultural wealth, silk, and arms manufacture, was threatened by extravagant taxation, in large part imposed to meet the cost of a virtually standing army.

The Sforza family

From 1463 to 1499 Milan also ruled Genoa. Bitter factional conflicts had, from the mid-14th century, eliminated Genoa as a political force and driven it to dependence on other powers. Yet, despite the advance of the Ottomans in the eastern Mediterranean, which threatened its colonies (Chios and Lesbos in the Aegean, Caffa in the Crimea, and Tana at the head of the Sea of Azov), its economy still prospered. With the support of the Bank of San Giorgio, which served as a state treasury, the city moved toward its 16th-century eminence as one of the great European financial centres. Genoese émigrés (such as, notably, Christopher and Bartholomew Columbus), discouraged now from settlement in the East, looked to new fields of enterprise in the Iberian Peninsula. By 1492 the city's bankers were dominant in Spain, particularly in Seville, and had already financed a considerable part of the exploration and settlement of the Canary Islands.

The first French invasion. Given that the rulers of both France and Spain had dynastic claims in Italy, it was predictable that with the end of the Hundred Years' War in 1453 and the conquest of Granada by Castile in 1492 both powers would at some time intervene in Italy and make the peninsula the battlefield of their conflicting ambitions. In the event, it was Italians who were to call the foreigner into Italy. King Ferdinand of Naples, angry that his grandson-in-law, Gian Galeazzo, duke of Milan, was excluded from power, threatened the regent, Ludovico il Moro. In reply, Ludovico successfully urged King Charles VIII of France to vindicate the claims of the royal house to Naples. Charles's response was, at first, stunningly effective. He crossed the Alps in early September 1494 and marched south. At Florence, Lorenzo's successor, his son Piero de' Medici, had declared in favour of Ferdinand. But the rapid advance of the French forces demoralized him, leading him to sue for peace in November. Discredited by this failure, Piero was forced to flee from the anger of his fellow Florentines. Charles entered Rome on the last day of the year and Naples on Feb. 22, 1495, having conquered "with the chalk of his billeting officers." Yet his triumph was short-lived. Alarmed at this sudden increase in French strength, Ludovico Sforza, the emperor Maximilian I, the pope, and King Ferdinand II of Aragon came together in the League of Venice in March 1495 to combat Charles's power. Faced by these forces, the king, leaving behind some of his troops in garrison, decided to return home. Crossing the Apennines by the Cisa Pass, he met the army of the league standing against his passage at Fornovo. After an indecisive battle, the French army broke through into Lombardy and passed back to France.

The League of Venice

Three years later, when Charles died, his campaign may have seemed merely a passing incident of no importance. Yet he had left behind a general ferment of change that profoundly weakened the peninsular states and rendered them almost powerless before a series of invasions that subjected them to domination by "barbarians" (as the Italians were pleased, now, to call non-Italians). Florence, humiliated by defeat and weakened by the establishment of a new government, struggled to regain control of those towns that had seized the occasion to throw off subjection. Naples, devasted by war, fell largely into the hands of Spanish troops in the pay of Ferdinand of Aragon. Ludovico Sforza, in Milan, was now deeply alarmed both by domestic unpopularity and the accession to the French throne of Louis XII, who as a member of the Orléans family laid a claim to be heir to the Visconti. Venice, characteristically emerging with spoils from the imbroglio (the Neapolitan ports of Otranto, Brindisi, and Trani), was looking for new triumphs, while Pope Alexander VI was reflecting on means by which the peace of Italy might be disturbed in the interest of his son, Cesare Borgia.

Savonarola. The French invasion and defeat and the exile of the Medici gave particular prominence within the new republican regime of Florence to a friar, Girolamo Savonarola. The son of a prominent physician, Savonarola had been born at Ferrara, had entered the Dominican order at Bologna at the age of 23, and had rapidly acquired fame as a theologian and preacher. In the years 1482–85 he had served in the convent of San Marco at Florence, returned there at the express wish of Lorenzo de' Medici, and became prior in 1491. In those years Savonarola preached conventional apocalyptic sermons warning of God's punishments that awaited Florentine sinners, including, notably, those guilty of evil in government.

Following the passage of Charles VIII's army, this message was transmuted into new forms. Drawing upon earlier Florentine mystical traditions, Savonarola now preached the doctrine that, in return for moral purification, Florence would soon become "the new Rome," enjoying power, dominion, and success in this world. This flattering teaching, the more acceptable in the circumstances of Florence's humiliation, brought a wide circle of personal adherents (the Piagnoni, or "Wailers," as their opponents called them), who enthusiastically backed Savonarola's campaigns (not in themselves untypical of revivalist movements of the age) against gambling, blasphemy, and illicit sex. From 1497 he organized bands of young men to go from house to house to persuade their inhabitants to surrender those worldly possessions to which they were particularly attached, such as dice, books, paintings, and elegant dresses. These "vanities" were then placed on a bonfire, and their destruction was solemnly dedicated to the Lord.

Bonfires of the vanities

In the controversy of 1495 as to what form of government should replace that of the Medici, Savonarola supported the party seeking the widest extension of popular participation. It is unlikely that Savonarola had any decisive influence on the political fortunes of the city; nonetheless, he came to be associated with the many failures of the government in those years and to be seen as an enemy both by the parties of the Bigi (looking for the return of the Medici) and the Arrabbiati (who hoped for a much more exclusive, less broadly based, republican government) as well as by the Compagnacci (those who resented the puritanical way of life now imposed on the city). In particular, he attracted enemies through the unflinching support he gave to an alliance with France, which, while isolating the commune in Italy, brought no response in loyalty from Charles VIII. This foreign-policy stance, as well as concern at his evangelical denunciation of the wickedness of the papacy, aroused the hostility of Pope Alexander VI. In June 1497 the friar was excommunicated and commanded to remain silent. Defying this decree, Savonarola resumed preaching early in 1498 and included in his sermons appeals for calling a general council to reform the church. Such defiance, combined with a certain revulsion against the unrelenting moral crusade, led in April 1498 to the secular government's turning against him. He was accused of heresy, tortured, and finally hanged and burned in the Piazza della Signoria (May 23).

Yet the contrast between the austerity of Savonarola's life and the licentiousness of the Borgia pope who condemned him, as well as the message of a particular destiny that the friar had held out to the Florentines, lingered in the minds of many, including some of the city's most distinguished citizens, and in the last Florentine republic, of 1527–30, the memory of his exalted prophecies was still to sustain those who resisted the Medici and the emperor Charles V.

THE EARLY ITALIAN RENAISSANCE

Against this political and economic background stands the cultural development of Italy in the 14th and 15th centuries. The term Italian Renaissance has not gone unchallenged; especially its meaning and boundaries have been the object of much controversy. From the 1340s the idea of "rebirth" was a commonplace in critical writing. Authors would speak of how, with Dante and Giotto, both poetry and painting had been "reborn," and in the following two centuries (normally in conjunction with another topos, "after a long delay") the same notion was often applied to other areas such as architecture, sculpture, and philosophy. In this period "rebirth" was always used in connection with some intellectual or artistic skill; it was not until the 19th century, with the French Romantic historian Jules Michelet and then, above all, the Swiss historian Jacob Burckhardt (whose *The Civilisation of the Renaissance in Italy* was first published in German at Basel, Switz., in 1860), that it became customary to write of the Renaissance as a period of time.

Definition of the term Renaissance

For Burckhardt this period consisted, broadly speaking, of the 15th century in Italy, a time and place in which "medieval" man became "modern" man. (The Italian of the quattrocento was "the firstborn among the sons of modern Europe.") No historian today would hold to that definition. Nonetheless, the term, redefined, still enjoys overwhelming assent, taken by some (such as Lauro Martines) to coincide with the life of the commune (and so going back to the 11th century), by others (such as Hans Baron) as springing from the ideological battles that accompanied the wars of Florence and Milan at the beginning of the 14th and 15th centuries; a majority consensus, however, still conceives of the Italian Renaissance as a period of cultural history having no very sharp chronological boundaries but stretching over the years circa 1340 to circa 1550.

Humanism. The early Renaissance can be characterized by two principal elements. Of these the first is humanism, to be thought of not in the ethical or antireligious sense in which the word has been used since the 19th century but rather as the intensive study of a revived classical antiquity. Humanism comprised an intense concern with the *studia humanitatis*—that is, grammar, rhetoric, history, poetry, and moral philosophy as read in classical Latin and, sometimes, Greek texts. As such it represented not a philosophical system but rather an educational program that largely excluded those subjects taught in the universities: logic, natural philosophy, metaphysics, astronomy, medicine, law, and theology.

The origins of humanism can be traced back to the Italy of the 1290s, in which one finds, in many cities, friends coming together informally to study the ancient world and attempting to reproduce something of the spirit of the Latin classics in their own writings. That the movement should have originated in Italy is not surprising. It was natural that Italians should look back to Rome, the more particularly in that the ruins of Roman civilization still stood about them. In addition, the study of the great corpus of Roman law in the universities of Padua and Bologna led easily to a wish to understand the society that had produced it. Yet even beyond that, in the secular world of the city-states, where intellectual life was dominated not by clerics but by lay literates, the secular civilization of the classical world made an irresistible appeal. It was not that the humanists were un-Christian, rather that their Christianity was a lay and, in some sense, secularized Christianity.

The movement was carried forward in the middle of the 14th century by two men, eminent both as humanists and for their roles in Italian and European literature:

Francesco Petrarca (Petrarch; 1304–74) and Giovanni Boccaccio (1313–75). It was consolidated at the end of the century, above all at Florence. Here in the 1390s the inspired teaching of the Byzantine Manuel Chrysoloras was to make the city the leading centre for the study of classical Greek in Europe, while Coluccio Salutati (1331–1406) and Leonardo Bruni (1370–1444), both of whom served for some time as chancellors of the republic, claimed that the disciplines of humanism were particularly suitable for the service of the state as studies appropriate to the "active life" of a republican citizen.

Thenceforth humanism was the dominant intellectual force in the peninsula (and later in much of Europe), influencing vernacular literature, the writing of history, art, education, and style of life. During the 15th century, Florentine Greek studies came, for the first time, to turn scholars from moral back to metaphysical philosophy. Marsilio Ficino (1433–99) translated all of Plato's writings, together with important Neoplatonic texts and the Greek mystical *Corpus Hermeticum*. From these sources he went on to develop his own philosophy of Christian Hermeticism, or Neoplatonism. Subsequently modified and developed by Giovanni Pico della Mirandola (1463–94), whose best-known essay bears the significant title *Oratio de hominis dignitate* (1486; *Oration on the Dignity of Man*), this philosophy, which argued that human beings could independently determine their own salvation by following the natural impulses of love and beauty, presented an immensely optimistic view of man and his place in the universe. It was to exercise a strong fascination, particularly over artists and poets, in the following hundred years.

The arts and intellectual life. Humanism does not, by itself, comprise the whole of the early Italian Renaissance, which should also be understood as a general intense efflorescence of all the arts and intellectual life. From the time of Dante and Giotto, through that of the great trio of Donatello, Brunelleschi, and Masaccio at the beginning of the 15th century, and on to the age of the High Renaissance, these years present a picture of extraordinary cultural power. In examining its social origins, it has been traditional to point to the economic wealth and early capitalist development of central and northern Italy. Certainly that development allowed the financing of patronage, advanced literacy, and in many ways offered a new way of looking at the world. Yet it should be emphasized that high culture developed unevenly throughout the peninsula—for instance, in this period it was insignificant in the great port and thriving economic centre of Genoa—so that the simple equating of wealth with culture cannot be sustained.

Social origins of Renaissance culture

It was, in fact, strong states (unlike Genoa) and the peculiar state system of Italy that lay behind most of the intense secular patronage and intellectual life in this period. In painting, sculpture, and architecture, the leading patrons were governments, and the motives behind the patron were a mixture of aesthetic response, civic pride, and propaganda. The communes made themselves responsible not simply for the *palazzi comunali*, or townhalls, and other communal buildings but also for the building, interior furbishing, and maintenance of their cathedrals and other principal churches (in these, sometimes specifically excluding any ecclesiastical participation in the work). In the same spirit, republics and *signorie* came to interest themselves in town planning—in the destruction and reconstruction of town sites, the regulations controlling building and use, and, by the appointment of conciliar committees, in the siting of new roads, squares, and fountains. At the same time, government involvement in the arts gave them an increasingly secular character. Political allegories and demands for identifiable portraits of lords or statesmen made new demands upon the artist and stimulated interest in the art of classical Rome, whose heir the communes claimed to be.

Whether in the republics or *signorie*, art had a major role as propaganda. Because Italy was divided into many states, political art was not centred, as in England or Scotland or France, at one court but was diffused throughout the peninsula. Because the states were in intense rivalry, art itself was enlisted in that rivalry. Thus the fragmentation of Italy, which made it so vulnerable to the foreigner from the last years of the 15th century, was also a cause of its supremacy in cultural life. At the same time, the papacy played its own part in this development—particularly from the mid-15th century, when Pope Nicholas V made the first full-scale alliance between the papacy and humanism, planning "majestic buildings, combining taste and beauty," to exalt the majesty of the Holy See.

(Jo.L.)

Early modern Italy (16th to 18th centuries)

FROM THE 1490S THROUGH THE 17TH-CENTURY CRISIS

The calamitous wars that convulsed the Italian peninsula for some four decades after the French invasion of 1494 are no longer seen by historians as the tragic aftermath of a lost world. Rather, they are perceived as a further elaboration and intensification of a violent age whose self-definition was transition. War reflected the wider European rivalries that made Italy a prize for plunder and a defensive bulwark against the Ottoman Turks, that led to the discoveries and conquests of the New World and to new contacts with Asia, and that erupted into open divisions over religious belief. Above all, war propelled all of Europe into a new economic and demographic expansion that was to shift the centre of power from the garden of Italy in the Mediterranean to northwestern Europe and its Atlantic world.

French and Spanish rivalries after 1494. The new political landscape after the 1494 invasion still reflected the contradictions and conflicts of the medieval political past. Rivalries of status, class, family, and neighbourhood continued unabated in the cities of both republics and principalities. Territorial states grew, and their urban capitals dominated neighbouring rural hinterlands even more than in previous decades. And, although independent action on the part of the Italian states was now seriously curtailed by powerful initiatives from the newly unified monarchies of France and Spain, such foreign intervention was entirely consistent with the policies of their medieval Angevin and Aragonese forebears.

French loss of Naples, gain of Milan. The French were not expelled from Naples. Charles VIII left Naples as freely in May 1495 as he had entered it a few months earlier. But an anti-French league led by Venetian and Spanish troops was needed to recover the kingdom for Ferdinand II of Naples (ruled 1495–96). When embattled French garrisons that had been left behind had their supply lines cut off by Spanish naval superiority, a preliminary armistice in 1497 ended the fighting.

The territorial designs of the Italian states led them to take advantage of the disequilibrium caused by the invasions. Venice, already more powerful than any of the other Italian states, gained the most. It occupied several important ports in Puglia with the intent of appropriating them, backed Pisa in its long though ultimately unsuccessful revolt against Florence (ending in 1509), and supported the conquest of Milan in 1499 by Louis XII (ruled 1498–1515), the new king of France, in exchange for Cremona and its hinterland.

Spanish acquisition of Naples. Louis XII had not given up French pretensions to Naples, and the acquisition of Milan strengthened his supply position. Powerful feudal lords, led by the pro-French princes of Sanseverino in Calabria, fomented dissension and weakened the already tenuous monarchy of Frederick I of Aragon (ruled 1496–1501) to the point that both the French and Spanish saw an opportunity to satisfy their ambitions. In the Treaty of Granada (1500) they agreed to invade and partition the kingdom between them into a northern French sphere of the Abruzzi and Campania (including the city of Naples), and a southern Spanish sphere of Calabria and Puglia. Yet the most wily diplomat of the age, Ferdinand II the Catholic of Aragon, hoped not only to forestall French ascendancy and outsmart Louis XII in Italy but also to assert his own claims as the legitimate heir to the Aragonese empire founded by Alfonso V the Magnanimous in 1442. In addition, he hoped to resist Ottoman advances that were threatening his possession of Sicily. In 1501 a French

The Treaty of Granada

and Spanish invasion divided the Kingdom of Naples according to plan, and Frederick of Aragon lived out his life in French exile together with his faithful servant, the great Neapolitan poet Jacopo Sannazaro. When hostilities broke out in Puglia in 1503 over the large revenues of the sheep customhouse at Foggia, Spanish forces under Gonzalo Fernández de Córdoba (the "Great Captain") outfought the French and laid claim to the entire Kingdom of Naples by the end of that year. During the next 30 years Naples spearheaded Spanish policy in Italy.

Tuscany and the papacy. During the first decade after the French invasions, Tuscany, the Romagna, and the Marche also underwent political upheavals. The Medici were expelled from Florence in 1494, and Savonarola's powerful sermons inspired a theocratic state. Tuscan cities that the French had liberated from Florentine rule continued their revolt. After Savonarola's execution in 1498, an oligarchic republic was created under the authority of Piero di Tommaso Soderini (ruled 1498–1512; elected gonfaloniere for life in 1502).

Meanwhile, Cesare Borgia, the natural son of Pope Alexander VI, attempted to carve out a dynastic state for himself in the Romagna and the Marche. As the model for political theorist Niccolò Machiavelli's prince, Cesare Borgia had prepared assiduously to seize power upon his father's death. But his plans were thwarted by bad fortune: at the very moment when decisive action was required, he himself was deathly ill. A college of cardinals caught between Spanish and French interests hastily elected a new pope, Pius III, who, however, died only 26 days later. His successor, Julius II (reigned 1503–13), had to win back by force of arms the territories in east-central Italy up to Bologna that Cesare Borgia had taken from the Papal States.

French victories in Lombardy. In order to reconquer the lost papal lands, Julius II organized an anti-Venetian alliance, the League of Cambrai (1508). All the great powers of Italy, along with those across the Alps—the Holy Roman Empire, France, and Spain—joined forces to defeat the Venetians at Agnadello (May 14, 1509). But dissension among the victorious allies, who were manipulated by skillful Venetian diplomacy, turned the alliance against France, because that kingdom now seemed to be the greatest power in Italy. A Holy League, organized in 1511 to curtail French power in Lombardy, restored the Medici in Florence in 1512 with the help of Spanish arms and allowed Venice to keep its old *terraferma* (mainland) empire (without its recent acquisitions in Lombardy, the Romagna, and Puglia). Nonetheless, Agnadello profoundly shook Venetian self-confidence and remained the turning point in the republic's imperial ambitions on mainland Italy.

The battle of Agnadello

At the same time Louis XII enjoyed his greatest triumphs, including the defeat of Julius II's Holy League at Ravenna (April 11, 1512). But, with the death of his brilliant general Gaston de Foix in that battle, the French suffered an irreparable loss. Further, in May 1512, 20,000 Swiss troops entered Italy on the papal side, and the French army was recalled to repel invasions of Navarre by the Spanish and of Normandy and Guyenne by the English. Francis I (ruled 1515–47), who succeeded his cousin and father-in-law, Louis XII, reopened hostilities in Italy. His army of 40,000 men defeated the Swiss at Marignano (Sept. 13–14, 1515), which allowed him to take Milan. The new pope, Leo X (reigned 1513–21), who was a Medici and a dependent of Spain, hurried to secure peace. Within the year, the new king of Spain, Charles I (ruled 1516–56), who had succeeded his maternal grandfather, Ferdinand II the Catholic, as coruler because of his mother's insanity, signed the peace of Noyon (Aug. 13, 1516), which gave Milan to France and confirmed Naples for Spain. The peace would not endure, however, as local Italian affairs became subordinated to the dynastic struggle between the young heirs to Habsburg and Valois (the ruling French dynasty) fortunes and to the Reformation movement that intertwined religion and politics into the 17th century.

The age of Charles V. Charles I, who was elected Holy Roman Emperor Charles V in 1519 upon the death of his paternal grandfather, Maximilian, aspired to universal monarchy over the far-flung territories he had inherited from Germany, the Netherlands, Italy, and Spain to the New World. The Piedmontese humanist Mercurino de Gattinara, Charles's chancellor from 1518 to 1530, fueled such ambitions, but the providential design for Charles to be a new Charlemagne collided with political realities. The revolt of the *comuneros* (1520–21), an uprising of a group of Spanish cities, was successfully quelled, securing Castile as the bedrock of his empire, but the opposition of Francis I of France, of Süleyman I the Magnificent (ruled 1520–66) of the Ottoman Empire, and of the Lutheran princes in Germany proved more intractable.

Early success in Italy, nevertheless, provided Charles with the most important base outside Spain for exercising his power. Imperial troops forced the French to retreat from Milan and restored the Sforza in 1522. When a refitted French army of 30,000 men retook Milan in 1524, the new Medici pope, Clement VII (reigned 1523–34), changed sides to become a French ally. But, at the most important battle of the Italian wars, fought at Pavia on February 24, 1525, the French were defeated and Francis I was captured. Soon after his release, he abrogated the Treaty of Madrid (January 1526), in which he had been forced, among other concessions, to abandon his Italian claims. He headed a new anti-Spanish alliance, the Holy League of Cognac (May 1526), which united France with the papacy, the Sforza, Florence, and Venice. With Charles's imperial army unpaid and no French forces in the field, some 12,000 troops, largely Lutheran infantry, marched south to Rome. On May 6, 1527, they attacked the city, forced the pope to take refuge in the Castel Sant'Angelo, and began to sack the city. The repercussions of this chastisement of the corrupt church were heard throughout Europe, and some scholars still date the end of the Renaissance in Italy to this event.

The Battle of Pavia

New warfare. New military technologies in siegecraft (cannon and bastion) and new techniques in open-field engagements (mixing pike and harquebus) not only transformed the nature of warfare but also threatened the order of a society still dominated by an aristocratic military caste. In the course of the Italian wars, the nonnoble infantry adopted tactical innovations that unseated the cavalry of heavily armoured nobility, which had dominated medieval warfare. Charles VIII's invading army employed the Swiss pike phalanx, whose moving squares of 6,000 men had already developed the ability to engage in offensive as well as defensive maneuvers. In the fighting against France for the Kingdom of Naples, Fernández de Córdoba first developed the Spanish *tercios,* more flexible units of 3,000 infantrymen using both pikes and harquebuses. Spanish military superiority eventually owed its success to the introduction in 1521 of the musket (an improved harquebus) and to the refinement of pike and musket tactics in the years preceding the Battle of Pavia. Such tactics dominated land warfare until the Battle of Rocroi in 1643.

The new social composition of the enlarged infantry as well as the need for large quantities of metal and the financial requirements for equipping and launching an army pushed military affairs further into royal hands, strengthening the growing power of the monarchy at the expense of the aristocracy. Commoners could forge a new relationship directly with royal authority. They could also, as in the case of the republics, create new images of citizenly power. In 1503 the Florentine republic, for example, planned two monumental mural paintings for the Great Council Hall of the Palazzo Vecchio (town hall) to be executed by two of the giants of High Renaissance art, Leonardo da Vinci and Michelangelo. The former's "Battle of Anghiari" and the latter's "Battle of Cascina," if completed, would have emphasized the strength and righteous rage of republican virtue and the necessity for citizens to be vigilant, challenging them to retake Pisa and subdue Tuscany during the republic's ongoing wars. Ludovico Ariosto, singing his epic poetry at the chivalric pro-French court of Ferrara, lamented the loss of glory, honour, valour, and courage to the "wretched and foul invention" of firearms. Even 20 years after the fact, when the diplomat and writer Baldassare Castiglione nostalgically portrayed the graceful court of Urbino of 1508 in *The Courtier* (1528), he did so in order to instruct courtiers and court ladies on how to adapt their roles to the changing times.

Spanish victory in Italy. In the immediate wake of the sack of Rome and the consequent disgrace of the Medici papacy, the Florentines expelled their Medici overlords. A French army under its general Odet de Foix Lautrec finally arrived in 1528, but Andrea Doria, a Genoese admiral and aristocrat whose galleys had formerly been in the service of the French, unexpectedly switched sides and became a staunch supporter of Charles V. Plague took Lautrec's life and decimated the French army, and in 1529 the pope was forced to make peace with Charles in the Treaty of Barcelona—as did Francis I in the Treaty of Cambrai. After almost 40 years of war, Italy submitted to Spanish pacification. Francis I renounced his claims in Italy, as well as in Artois and Flanders. The last Sforza was restored in Milan with the provision that the duchy would revert to Spain upon his death. Venice lost its mainland conquests. The Papal States were restored, and in 1530 the pope crowned Charles V emperor and king of Italy, even making vague promises to call a council to address the Protestant schism and reform the church. In exchange, the Medici were reinstated in Florence through the aid of Spanish arms.

Spanish pacification of Italy

Italy remained subject to sporadic French incursions into Savoy in 1536–38 and 1542–44 during a third and fourth Habsburg-Valois war, and Spain's Italian possessions were increasingly taxed to support Charles's continual campaigns; however, for the remainder of his reign, Charles's armies fought the French, the Ottomans, and the Protestant princes outside Italy. Notable for Italy was Charles V's capture of Tunis in 1535 and his glorious march up the Italian peninsula in 1536 to confirm his personal rule. But the Ottomans formally allied themselves with France against the Habsburgs thereafter, defeated an allied fleet at Prevesa, retook Tunis in 1538, and stepped up their assault on the Venetian empire in the Mediterranean. With the eventual failure of Charles's attempts to secure Germany, his great continental empire was divided. Italy became a part of the Spanish Habsburg inheritance of his son, Philip II (ruled 1556–98), and, after the Spanish victory over the French at St. Quentin (1557), the Peace of Cateau-Cambrésis (1559) officially confirmed the era of Spanish domination that had existed in Italy since 1530.

Spanish Italy. Spain thus established complete hegemony over all the Italian states except Venice, which alone maintained its independence. Several Italian states were ruled directly, while others remained Spanish dependents. Naples, Sicily, Sardinia, and Milan came under Spanish rule through the kingdom of Aragon and owed their allegiance to the sovereign according to their own laws and traditions. Their foreign policy interests were subordinated to the imperial designs of Spain, which also appointed their chief officers (viceroys in Naples, Palermo, and Cagliari; a governor in Milan) and administered their internal affairs through local councils. From the beginning of Philip II's reign, Italian affairs, which had originally been administered by the Council of Aragon, were coordinated by a Council of Italy in Madrid. At this council, the three major states—Naples, Sicily, and Milan—were each represented by two regents, one Castilian and one native. Sardinia remained a dependency of Aragon. The king, however, continued to receive and be responsive to embassies sent by various groups outside official channels until the Spanish Habsburg line died out in 1700.

A vitriolic anti-Spanish polemic has long dominated the historiography of early modern Italy. It accuses Spanish rule of an authoritarianism closed to new ideas and innovation, of presiding over an empty formalism in literary expression, and of promoting *spagnolismo,* an exaggerated and ostentatious pomp—all perceived as the fruits of a decadent, backward-looking colonial domination. Faulting Spain for trying to integrate Italy within its absolutist and imperial program or blaming Italy's 17th-century decline on Spanish social and economic policies has served nationalistic fervour since the 16th century, but it has missed both the benefits of Spanish rule to Italian peace and security and the causes for crisis in 17th-century Italy. To understand the latter, one must examine the internal conflicts and economic impediments that existed within the Italian states themselves rather than look to an absentee Spanish scapegoat. And, above all, early modern Italy must be understood in a wider European context and in relation to the economic shifts wrought by the new Asian and American trade. The touchstone for modern scholarship is Fernand Braudel's *The Mediterranean and the Mediterranean World in the Age of Philip II* (1949), which continues to inspire and challenge research into Philip II's empire and beyond.

Pedro de Toledo

The Kingdom of Naples. Pedro de Toledo (viceroy 1532–53) reorganized the Kingdom of Naples and placed it firmly within the Spanish monarchical orbit dominated by Castile. Within the kingdom, he oversaw the eradication of the pro-French barons and attempted to install centralized, absolutist policies. Within the city, he developed new residential quarters and strengthened Spanish defenses against outside attack. He enjoyed unparalleled personal prestige; his daughter Eleonora was married to Cosimo I the Great, the Medici duke of Tuscany, in 1539. But his power had limits, as was shown by the successful Neapolitan opposition to the introduction of the Inquisition in 1547. Pedro's policy was governed by the principle of "divide and conquer," which played upon rampant inequalities between the barons and the people and between the capital and the countryside.

The most important ruling body in the kingdom was the Collateral Council, composed of five regents presided over by the viceroy, with a judicial council and a financial council exercising their respective competencies at its side. A new elite of lawyers, a "nobility of the robe," began to emerge, sustaining the Spanish regime with its indispensable bureaucratic services. The Neapolitan parliament, which consisted of representatives of the city districts (*seggi*), of the feudal nobility, and of royally owned towns, had only two functions—to authorize taxes and to request rights and privileges from the king in exchange—but this body was suspended in 1642.

In the capital the town council, which seated representatives of the city's five noble *seggi* and of a citywide commoners' *seggio,* emerged as the most important institution of municipal government. The most pressing problem facing the city administration was the provision of food. Naples had grown to 250,000 inhabitants by 1600, which ranked it first in population among the cities of western Europe.

In the countryside, where some 90 percent of the population still lived, the feudal nobility retained strong social and economic control. The Spanish government's bureaucratic administration did attempt to break the barons' political stronghold and to limit the worst abuses, but success depended upon a healthy economy and an emerging middle class, both of which began to falter after 1585. The 12 provinces of the kingdom remained atomized, and their unarticulated markets were often attached to the trading networks of foreign states such as Venice or Genoa rather than integrated to form a national market within the kingdom itself.

The kingdom of Sicily. Sicily's administration existed apart from that of the mainland since 1282, when the island had revolted against Angevin rule and come under the Aragonese crown. In the 16th century Sicily remained the cornerstone of the Spanish Mediterranean policy against the Ottomans, and its agricultural products continued to be the staple of long-distance trade.

Spanish policy in Sicily

As in Naples, Spanish policy in Sicily attempted to modify traditional baronial abuses. Spain allowed the barons considerable autonomy over their large agrarian estates, including the exploitation of their tenant farmers, but it prevented open feuds between barons and eroded their political power by excluding them from offices in the central government. Two local councils, one in judicial affairs and the other in public finance and administration, centralized Spanish government from the reign of Charles V. Parliament and the Inquisition competed for power with the viceroy. Parliament, which was composed of three branches—clergy, nobility, and royal towns and districts—voted ordinary and special taxes, but its short and infrequent sessions prevented it from mounting sustained opposition to Spanish policies. The Inquisition, on the other hand, was completely independent of the viceroy and often challenged his jurisdiction, but it received royal backing only in purely religious disputes. Above all, Spain played internal rivalries and sectional interests against one another for its own advantage. Constant struggles weakened all parties, and the numerous autonomous authorities held civil government in such check that it became immobilized and unable to make important decisions.

Sardinia. Sardinia's links to the king of Aragon dated from the 14th century. Long-standing assimilation to Spanish culture had reinforced the patriarchal structure of the local feudal nobility, whose chief source of wealth was sheep raising. As in Naples and Sicily, the Spanish introduced little change into government, preferring instead to support a feudal-monarchist regime. The viceroy was often a Sardinian, the native parliament had three branches, and international politics separated Sardinia from Italian affairs.

The duchy of Milan. When Francesco II Sforza died childless in 1535, Milan devolved to Charles V and was administered by a Spanish governor, who maintained traditional institutions. The duchy consisted of nine provinces, each dominated by a small group of families resident in their provincial capitals. Central administration from Milan rested primarily with the Senate, a judicial and legislative body that maintained its authority under Spanish rule despite inevitable confrontations with the governor. Official Spanish policy aimed at maintaining an equilibrium between centralization and home rule.

Two institutional changes, nevertheless, had significant effects upon Lombard society. First, by 1584 the membership of the Senate was reduced from 28 to 15 as well as altered in its composition; whereas half of its members had been feudal lords and high-ranking clerics, they now were all professional lawyers. As in Naples, the nobility of the robe, who in Milan were lawyers drawn from the urban patriciate, grew at the expense of the old feudal nobility and formed an essential alliance with the Spanish crown. Second, tax reforms aimed at marshaling Milanese resources for the Spanish wars affected Lombard society, not only in equalizing the tax burden but also in redistributing power between city and countryside. Merchants, who had previously been tax-exempt, found their wealth (based on annual gross sales) taxed after 1594, and landowners not residing in cities, who had previously been taxed far above city dwellers, benefited from a new assessment system set by elected bodies of rural residents after 1561. Unexpected long-term effects of these policies were felt in the 17th century when economic interests were able to regroup and find a foothold in the countryside.

Milan's strategic importance as the gateway to Italy remained a tenet of Spain's imperial design, and, with war and revolt north of the Alps, Milan served as a critical staging area for men and supplies on the "Spanish road" from Genoa to Lombardy and from there through the Alpine passes to the Rhineland. From the Revolt of the Netherlands (1567–79) through the Thirty Years' War (1618–48), Milan was a focal point of Spanish military preparation.

Charles Cardinal Borromeo

The Roman Catholic church had unusual influence and autonomy in Milan. Charles Cardinal Borromeo, member of a rich noble family of Milan and nephew of Pope Pius IV (reigned 1559–65), resided in his diocese after 1565 as the model Counter-Reformation bishop. He instituted seminaries, diocesan synods, and provincial councils, personally visited some 800 parishes, watched over the spiritual needs of monasteries, convents, and lay confraternities, fought heresy, and supported relief of the poor. Moreover, under his rule the Lombard church enjoyed unusual freedom of action and special privileges in furthering Catholic reform.

Principates and oligarchic republics. Spanish hegemony in Italy extended beyond the states under its direct control. The rulers of Savoy and Tuscany owed their titles to Spain, Genoa acted as Spain's chief banker, the papacy depended heavily on the Spanish monarchy in the age of the Counter-Reformation, and even independent Venice needed Spanish aid in protecting the Mediterranean from further erosion by the Turks. A number of minor states were prevented by their small size from having much political influence; these included the republic of Lucca as well as several duchies that remained under the control of local noble families—the duchies of Modena, Reggio, and Ferrara under the Este family; the duchy of Mantua and Montferrat under the Gonzagas, and the duchy of Parma and Piacenza under the Farnese. These states, too, enjoyed the enforced Spanish peace within Italy and benefited from the security against foreign invasion. Their nobility intermarried with the Spanish aristocracy and absorbed Spanish culture.

The duchy of Savoy. During the Italian wars, France and Spain had occupied Savoy, a duchy between France and Spanish Lombardy. Allied with the victorious Spanish at the battle of St. Quentin (1557), its legitimate heir, Duke Emmanuel Philibert (ruled 1559–80), recovered his state with the Peace of Cateau-Cambrésis (1559) and began to rebuild and strengthen it. He transfered the capital across the Alps from Chambéry to Turin, which grew as a fortified and planned city. He limited the power of numerous localities and centralized state finances. Increased taxes and economic recovery allowed him to maintain a small but disciplined standing army, which became the basis of Piedmontese military power. His son Charles Emmanuel I (ruled 1580–1630) followed an expansionist policy with varying success. In 1589 he failed to take Geneva, and in 1601 he ceded some territory to France in exchange for the marquessate of Saluzzo. He also engaged in debilitating wars in an unsuccessful quest to take Montferrat.

The duchy of Tuscany. When Spanish arms restored the Medici to Florence in 1530, they recognized the title duke of Tuscany. After the assassination of the first duke, Alessandro, in 1537, Cosimo I (ruled 1537–74) succeeded him and developed a strong absolutist state. As a Spanish ally, Cosimo fought Siena (1552–55) and annexed it in 1557. The Spanish, however, retained five strategically important seaports, the Stato dei Presidi ("State of the Garrisons"), which were administered by Spanish Naples.

In 1569 Cosimo received the title grand duke of Tuscany. His sons Francis I (ruled 1574–87) and Ferdinand I (ruled 1587–1609) succeeded him, and the latter enlarged the free port of Livorno. In the early modern period, the city of Florence had a population amounting to only about one-half of its medieval one, and it receded from the international scene, becoming the capital of a provincial court.

The republic of Genoa. In 1528 Andrea Doria initiated a constitutional reform by which nobles loyal to him gained power. Factionalism continued, however, especially between the "old" and "new" nobility. When serious disorders erupted in 1575, the old nobility abandoned the city, and a popular faction took their place beside the new nobility. A compromise mediated by Spain and the papacy averted civil war by reconstituting the ruling class; wealth replaced status as the basis of social stratification and political alliance.

Andrea Doria's support of Charles V bolstered Spain's naval profile in the western Mediterranean. Genoa continued its control over Corsica through its central bank, the Bank of San Giorgio. Genoese bankers, who had extended their familial-business interests from Naples to Seville, replaced the German Fuggers as the primary financiers of the Spanish empire. At home, nobles invested in landed property and city residences, while silk manufacturing employed a large percentage of the Genoese working class.

Contraction of the Venetian empire

The republic of Venice. Defeat at Agnadello in 1509, followed by pressure from the Spanish Habsburgs in Lombardy and the Austrian Habsburgs to the north of the republic, limited Venice's Italian mainland empire. In addition, Ottoman expansion in the eastern Mediterranean disrupted Venice's trade in the Levant and chipped away at its overseas empire: lost were important ports of call in Albania and Greece in 1503, the Aegean islands north of Crete in 1540, Cyprus in 1571, and Crete itself in 1669. At the same time, Portuguese trade to Asia after 1498 and the rise of the Dutch city of Antwerp as an entrepôt for the distribution of goods to northern Europe seriously challenged Venice's trading monopoly. No longer the most powerful state in Italy, Venice still enjoyed internal cohesion, an extremely effective diplomatic corps, and a strong fleet to navigate an independent policy between Spain and the papacy.

Before the plague of 1576, Venice's population had risen to 180,000, with a patriciate of under 5 percent. A strong oligarchic tendency during the 16th century reinforced the power of the Council of Ten over the Senate, and the cleavage between rich and poor nobles widened. After 1583, however, the old nobility lost its bid to monopolize politics, and the Senate recovered power, which it applied to a more independent foreign policy. Textile manufacturing remained the most important trade until the precipitous decline of the woolen industry in the early 17th century. Venice's population stabilized at about 150,000.

Paolo Sarpi

In 1606 a papal interdict condemned Venice for refusing to repeal several laws limiting the church's traditional rights and for trying two priests in civil rather than ecclesiastical courts. Paolo Sarpi, the republic's state theologian, mounted an effective defense by arguing for state sovereignty in temporal affairs. The dispute ended in a compromise, mediated by France and Spain. Sarpi's *The History of the Council of Trent* (1619) later indicted the pope for usurping ecclesiastical authority and for manipulating the reform council to reinforce his power.

The Papal States. The papacy was engaged in often flamboyant political maneuvers, especially during the reign of Julius II (1503–13), and in the architectural and intellectual renewal of Rome. Save for the brief reign of the last non-Italian pope before the 20th century, Adrian VI (reigned 1522–23), the papacy failed to respond to the spiritual crisis of the day. However, a predisposition for a religious revival, or Catholic Reformation, was fostered by the Christian humanist Erasmus of Rotterdam's biblical philosophy of Christ, by prophetic and apocalyptic interpretations of the Italian wars, and by an awareness of long-standing clerical abuses. Yet serious attempts at reform from above were initiated only with Pope Paul III (reigned 1534–49). In 1536 he appointed a reform commission, which produced the important blueprint *Consilium de emendanda ecclesia* ("Project for the Reform of the Church"), and in 1537 he made the first attempt at convoking a reform council. By the 1540s, however, hopes for reunification of Catholics and Protestants had foundered. A true Counter-Reformation—that is, the Roman Catholic church's conscious fight against Protestantism—began to take shape with papal approval of the Jesuit order in 1540 and with the creation of the Holy Office of the Inquisition in 1542. New religious orders such as the Theatines (1524), the Capuchins (1528), and the Jesuits (1534) provided the backbone of the new class of religious leaders, although the apostasy to Calvinism in 1542 of Bernardino Ochino, vicar-general of the Capuchins, was a serious setback. The Jesuits' educational program, above all, began to prepare laymen of high social status for leadership roles. The Council of Trent (1545–63) uncompromisingly defined Catholic dogma and outlined a program for disciplinary reform and administrative centralization. The *Index librorum prohibitorum* ("Index of Forbidden Books"; 1559), a list of books condemned by the Roman Catholic church as being pernicious to faith and morals, was compiled by a censorship board that limited orthodox expression to a narrowly controlled range of inquiry.

Council of Trent

In politics the papacy found itself dependent on Spain yet desirous of finding an alternative to Spanish domination in Italy. Although ecclesiastical reform occupied most of the church's energies, Pope Pius V (reigned 1566–72) promoted the Holy League, which checked Ottoman expansion into the western Mediterranean by defeating the Ottoman fleet at Lepanto (1571). Under Pope Gregory XIII (reigned 1572–85) the Julian calendar was reformed into the modern Gregorian calendar. Pope Sixtus V (reigned 1585–90) launched a Catholic missionary counteroffensive in central Europe and reorganized the Roman Curia. He, along with Clement VIII (reigned 1592–1605), also patronized the urban development and new artistic flowering in Rome that culminated in the Baroque creations of Lorenzo Bernini and the architect Francesco Borromini. These two popes also fought rural banditry and brought Ferrara, Urbino, and Castro back under direct rule.

Culture and society. Cities and courts spawned the high culture of late Renaissance Italy. Ranging from Pietro Aretino's merciless lampoons of the scandalous lives of the princes of the church in Renaissance Rome to the mysticism and Christocentric piety embraced by the intellectual circle surrounding the Spanish humanist Juan de Valdés in Naples, Italian culture in the 16th century defined itself for or against the church. Machiavelli, in a famous chapter of *Discorsi sopra la prima deca di tito Livio* (written 1513–19; "Discourses on the First Ten Books of Livy"), argued that the church was the cause of Italian ills because it had lost its religious moorings and had kept the Italians politically divided. A rigid Counter-Reformation orthodoxy, however, condemned some of Italy's most brilliant intellectuals—philosophers and scientists such as Giordano Bruno, who was burned as a heretic in 1600, Tommaso Campanella, who was imprisoned in 1599 for 27 years, and Galileo Galilei, who was forced to recant his Copernican beliefs and was placed under permanent house arrest in 1633.

At the same time, however, Italy was at the forefront of a movement that fostered scientific exchange by establishing scientific academies—the Roman Accademia dei Lincei (founded in 1603), the Florentine Accademia del Cimento (1657), and the Neapolitan Accademia degli Investiganti (1665). In fields such as drama (both tragedy and comedy), music (both religious and secular), art history, rhetoric, and political theory, formative roles were played by Italians of the late Renaissance—the poets Torquato Tasso and Giambattista Marino, the composers Giovanni Pierluigi da Palestrina and Claudio Monteverdi, the artist and art historian Giorgio Vasari, and the political theorist and statesman Giovanni Botero, to name just a few. As research into Italy's forgotten centuries continues, the old nationalistic historiography that argues for a decline into a dark age and a disjunction between the period of the Italian Renaissance and that of the 18th-century reformers dissolves.

Society and economy. The expanding demographic and economic base of Italy provided the wherewithal for the political and cultural programs of the 16th century. From the mid-15th-century demographic low point after the 1347–48 plague, Italy, along with the rest of western Europe, recovered dramatically. Between 1400 and 1600 the Italian population nearly doubled, increasing from 7 million to *c.* 13 million, and prices rose sharply, with cereal prices tripling and quadrupling. Increased demand, the increased supply of money from the silver of the New World, and profligate military expenditures fueled high inflation. Italy's most distinctive feature was its highly urbanized life. In 1550, 30 cities—more than in any other region in the West—had populations of more than 10,000.

Urbanization of Italian life

The rural population, nevertheless, still accounted for almost 88 percent of the total population, and, given the relative parity in birth and death rates, cities grew primarily as a result of rural emigration. Wheat and wool were the chief agricultural products, and the spread of capitalist agriculture in the 16th century was an important ingredient in the transition from feudalism to capitalism. Textile production of both woolens and silks continued to be the major industry in the cities, but the precocious economic development of Italy in manufacturing, trade, and finance came to a crashing halt during the dislocations of the 17th century.

The economic recovery of the second half of the 16th century challenged the traditional hierarchical ordering of society. Nobility and clergy, the two most identifiable groups, did not lose their status but slowly changed character. With the demise of old families and the rise of a new nobility based on wealth and public service, social mobility in the cities put the old aristocracy on the defensive until it was able to forge new alliances with the ruling princes and the bourgeois bankers and merchants. At the same time, demographic growth and a yawning gap between wages and prices threatened to create an even larger disparity between rich and poor. In good times the lower classes could provide a new labour market, fueling industrial production; in economically bad times, however, sickness, unemployment, and the rising price of

The lower social classes

bread could drive them down into the vortex of poverty and even push them to the point of rebellion.

The 17th-century crisis. The economic boom of the late 16th century began to stall throughout Europe. The first signs of hardship appeared in Italy after 1585, and famine persisted through the 1590s. New waves of plague struck northern Italy and Tuscany in 1630–31 and southern Italy, Lazio, and Genoa in 1656–57, with population losses between one-fourth and one-fifth, respectively. The large cities of Milan, Naples, and Genoa lost as much as half of their population. In addition, war in northern Europe after 1618 and in the Middle East between the Ottomans and the Iranians from 1623–39 disrupted Italy's important export markets; war between Spanish, German, French, and Piedmontese forces moved to Italy between 1628 and 1659; and social conflicts within the Spanish states contributed to the decline of Italy relative to northwestern Europe.

Both agricultural production and urban industries entered into crisis in the decade 1611–20, reaching their low point about 1650. In the south, extensive wheat monoculture exhausted the soil and led to deforestation and soil erosion. Further, noble owners drained off profits for expenditures on urban luxuries, and indebtedness placed commercial grain farmers at greater risk as grain prices fell in the 17th century. In the north, intensive agriculture supported the numerous large cities, but overexpansion onto unproductive land, soil depletion, and the loss of credit pushed the region to the limits of what the population could support. In the cities, wool manufacturing fell by 50 percent in the 1620s and all but disappeared thereafter, although silk production held its own. Commercial and banking activities, once the fastest-growing industries, now constricted, and foreign imports braked further development at home. Italy's early industrial lead lost to increased competition from northwestern Europe as new products at lower prices replaced the traditional ones in the Italian markets. The Italian guilds' opposition to technological and organizational change, higher taxes, and higher labour costs prevented the adaptability required to surmount the short-term crisis, which instead turned into a long-term structural realignment. Only in Lombardy was there a successful shift to the putting-out system, which transfered urban industries to the countryside.

The economic involution reinforced the social hierarchy, favoured investment in landed property and rents over commerce and industry, and reinvigorated noble pretensions. With capital shifted from the manufacturing and service sectors to agricultural production of cash crops such as olive oil, wine, and raw silk, the number of skilled urban craftsmen and merchants decreased while that of illiterate peasants increased, and landed noble power intensified. The church reasserted itself in every aspect of social life, from land ownership to ecclesiastical organization, from the defense of orthodoxy and the culture of the Council of Trent to the education of the ruling class. As the economic crisis deepened, middling ranks lost out, and social stratification between rich and poor rigidified.

In the political sphere, Spain's involvement in the Thirty Years' War (1618–48) drained Italy. As Spain declined, it dragged its Italian realms down with it. Revolts broke out in Palermo and Naples in 1647. In Naples a revolt of July 7 is mistakenly identified as a plebeian rebellion bearing the name of a young fishmonger, Masaniello, although he was murdered within 10 days and had actually been a tool of bourgeois elements seeking greater political power in the city. The uprising spread to the countryside, established a republic seeking French protection, and assumed the character of an open rebellion against Spain and native feudal lords. Internal dissension and the arrival of the Spanish fleet brought an end to the revolt by April 1648. The social and economic crisis deepened in Naples after the failure of the revolt and a recurrence of the plague in 1656. Lost was any hope of an antifeudal alliance between the middle classes and the urban proletariat or rural masses. Paradoxically, renewed Spanish reliance on the nobility of the robe fostered the very class that was to lead the cultural renewal which made Naples one of the intellectual centres of 18th-century Italy.

The Masaniello revolt

REFORM AND ENLIGHTENMENT IN THE 18TH CENTURY

Upon the death of the last Spanish Habsburg, Charles II (ruled 1665–1700), fighting for the Spanish spoils consumed the European powers in the War of the Spanish Succession (1701–14). The treaties of Utrecht (1713) and Rastatt (1714) confirmed a new pattern of state relations in Italy between Austrian Habsburgs, Spanish Bourbons (with Bourbon France always in the background), and the independent states. After complicated military and diplomatic maneuvers, this pattern eventually stabilized into a long-term equilibrium. Naples, Sardinia, and Milan (which had incorporated Mantua after the last Gonzaga had sold it to Louis XIV in 1701) passed to the Austrian Habsburgs. Sicily went to Victor Amadeus II, duke of Savoy, who assumed the title of king of Sicily; after renewed Spanish hostilities, however, Victor Amadeus was forced to exchange Sicily for Sardinia in the Treaty of The Hague (1720). In 1734, as part of the War of the Polish Succession, Charles, son of the Bourbon Philip V of Spain, conquered the Kingdom of Naples. After the Medici dynasty in Tuscany died out in 1737, Francis Stephen, Duke de Lorraine, husband to Maria Theresa of Austria and Holy Roman emperor after 1745, ruled as grand duke of Tuscany from Vienna under a regency. And in 1748, after the War of the Austrian Succession (1740–48), Austria regained Milan (which had been lost more than once in the preceding years), and Charles of Bourbon's brother Philip was assigned Parma and Piacenza.

Society and economy. A slow economic recovery began in Italy in the mid-1680s, but it remained weak into the early 18th century. A slump in the 1730s was surmounted by strong economic growth during the mid-century, until the famines of 1763–67 highlighted the weakness and inefficiency of government policies. Regional differences in Italy's agricultural structure led to even greater divergences between north and south. Whereas some northern urban industries found refuge in smaller centres and rural settings, the south became almost exclusively an agricultural producer. Overall, Italy's foreign trade decreased and shifted from producing and exporting industrial goods to importing finished industrial products and exporting raw materials and semifinished goods. At the same time, the Italian national market also contracted, and tightening social and institutional constraints further closed off productive and mercantile opportunities. While Italy's population between 1700 and 1800 rose by about one-third, to 18 million, that of the rest of Europe grew at twice that rate. Neither an agrarian nor an industrial revolution could be foreseen for 18th-century Italy.

The noble classes retained hegemonic control of politics and economics, monopolizing land ownership and manipulating legal and political institutions in the towns to maintain their position. Tensions and conflicts arose from time to time between the central authority of the absolutist states and the nobility, between the rich bourgeoisie or professional classes and the nobility, and among the nobles themselves, but they were managed or controlled, as the old social hierarchy blocked, worked out compromises with, and co-opted the new aspirant groups. In the north, especially in the republican states, city oligarchies resisted erosion of their power and privileges. In sharp contrast, the urban masses, and especially the growing rural population, suffered a deterioration of their social and economic positions as they faced the difficulties of daily life.

The position of the nobility

Political thought and early attempts at reform. By the beginning of the 18th century, a new cultural climate opened Italy to a wide range of European ideas—especially the philosophical thought of René Descartes, Pierre Gassendi, Benedict de Spinoza, Pierre Bayle, Thomas Hobbes, John Locke, Sir Isaac Newton, and Hugo Grotius. With it new cultural institutions came to the fore. The Academy of Arcadia, founded in Rome in 1690, is exemplary of the channeling of energies for rationalism and innovation. Among its more famous members, Gian Vincenzo Gravina, Ludovico Antonio Muratori, and Giambattista Vico distinguished themselves as literati by launching juridical, historical, aesthetic, and "scientific" critiques of society. Vico's *New Science* (1725; 3rd ed. 1744), eventually the most enduring work produced by

Vico and Muratori

this group, found tepid reception in its own day, and the author's ideas on a universal philosophy of history were integrated into Enlightened thought only in the 1770s. Paolo Mattia Doria (1662?–1746) and the Medinaceli Academy in Naples also employed historical inquiry to explore remedies for society's present ills. Doria revived the idea of a Platonic republicanism of philosophic magistrates, in which an anti-Enlightenment Catholicism would become a kind of civil religion. In Naples he led his group of self-styled "ancients" against the scientific "moderns" led by the Neapolitan diplomat Celestino Galiani and Bartolomeo Intieri, a Florentine factor in Naples who provided a link to Tuscan intellectual circles. Such networks of intellectual exchange between the cities of Italy and between Italy and the broader cosmopolitan centres of 18th-century Europe had deep roots in the ministerial class fostered in Spanish Italy from the early 16th century.

The political and cultural roles of the church—in particular, the supranational character of the papacy, the immunities of clerics from the state's legal and fiscal apparatus, the church's intolerance and intransigence in theological and institutional matters, as well as its wealth and property—constituted the central problems in the reform schemes of Italy's nascent intellectual movement. Pietro Giannone (1676–1748), a Neapolitan jurist, provided the most incisive breakthrough by employing a jurisdictional, historical method to oppose church abuse of power and to break the church's stranglehold on the state. Francesco Scipione, Marchese di Maffei (1675–1755), and Muratori (1672–1750), offered probably the strongest rationale for a moderate Enlightened Catholic reform by seeking to reconcile politics with morality and religion. Muratori's *Della pubblica felicità* (1749; "On Public Happiness") reached Bourbon audiences in French and Spanish translations and was probably read in the Austrian Habsburg realms by Maria Theresa herself.

Politics and Enlightened despotism. By the mid-18th century the first stirring of reform was produced by the confluence of economic recovery, Muratori's program for Enlightened Catholicism, and a renewed interest in natural science, political economy, and agronomy. The dynasties installed after the wars of succession—the Habsburg-Lorraine in Lombardy and Tuscany and the Bourbon in Naples—led the way.

Lombardy. A first wave of reforms under Maria Theresa came to Milan in the early 1740s. It was linked with the Genoese patrician Gian Luca Pallavicini, who first began to prepare them as a minister after 1743 and then implement them as governor after 1750. Among the major reforms, government administration was reorganized and offices were no longer sold, state finances were reordered and a public bank was founded, and, most important, a new cadastral survey—begun in 1718 but interrupted in 1733—was placed under the direction of the Florentine jurist Pompeo Neri in 1749. The Theresian cadastral survey took effect in 1760. By applying objective principles of fiscal justice and administrative rationalization, the new method of registering the ownership and value of property not only revamped Lombardy's fiscal system but also had the effect of improving agriculture, increasing productivity, and centralizing control of revenues in impartial hands.

Gian Luca Pallavicini

A second wave of reforms, in the 1760s, was also orchestrated from Vienna, where a Department of Italy had overseen Lombard affairs since 1757. The arrival in 1759 of another imperial official, Carlo, Count di Firmian of Trent, initiated wide-ranging change. Firmian completed the earlier reforms in political administration, in the judicial system, in ecclesiastical relations, and in educational policy. But strong opposition from diverse social groups defending traditional rights and privileges weakened the reform movement. In 1761–62, however, an important group of young noblemen engaging in reformist thought formed around Pietro Verri (1728–97) and took the name of his militant journal, *Il caffè* (published 1764–66; "The Coffeehouse"). The circle's best-known work, Cesare Beccaria's *Dei delitti e delle pene* (1764; *On Crimes and Punishments*), castigated torture and capital punishment as symptoms of the injustice and inequality inherent in the society of the old regime.

Pietro Verri and Cesare Beccaria

Joseph II (ruled 1765–90) promoted a new wave of reforms after 1770 that gained strength when he became the sole ruler after Maria Theresa's death in 1780. The old system of public administration and magistratures came under attack and was abolished by 1786. In the 1770s and '80s the reform policies of "Josephism" succeeded in suppressing all the chief political and judicial bodies of the Lombard aristocracy and in establishing modern ones in their place. Provincial intendants were appointed, and the church's power in the state was reduced. Educational reform established popular elementary schools as well as new disciplines at the Palatine School of Milan and the University of Pavia.

Such reforms, however, proved to have little long-term cultural or social consequences. Opposition from nobles, local administrators, feudal landlords, magistrates, clergy, and even Enlightenment intellectuals, who feared Joseph's new authoritarianism, undermined the reforms. Leopold II (ruled 1790–92), who gave up his Tuscan domain, which he had ruled as Duke Peter Leopold, to succeed his brother Joseph, could not overcome their resistance, now strengthened by the forces unleashed with the French Revolution. Francis II, who succeeded Leopold II in 1792, was absorbed in matters of war and politics that led to French victory in Italy and occupation of Lombardy in 1796.

Tuscany. The regency of Francis Stephen in Tuscany (1737–65) was distinguished by his chief representative for 20 years, Emmanuel, Count de Richecourt, who followed the main lines of Habsburg policy in Lombardy. Local aristocratic divisions, the privileged position of Florence, the Tuscan capital, and the corruption and private enrichment of public officials came under scrutiny. Reforms were aimed at restoring revenues, reorganizing magistratures, controlling the old nobility, and moderating the influence of the church. Pompeo Neri, who was recalled from Milan to Florence in 1758, addressed the problems of economic scarcity by advocating the free trade of cereals in order to provide incentives to agricultural production.

Reforms in Tuscany

Physiocratic solutions to economic problems—that is, solutions based on laissez-faire economics and on the belief that land is the source of all wealth—characterized Tuscany under the leadership of Peter Leopold (ruled 1765–90). The Accademia dei Georgofili, founded in 1753, exercised a wide influence on a range of issues touching on agrarian reform. In 1767 the free trade in grain was confirmed by law, in 1771 art and craft guilds were suppressed, and in 1781 all internal customs dues were eliminated. Peter Leopold planned to redistribute land owned by the church and state to a new class of independent small farmers. These, in turn, would form a genuine foundation for a new kind of polity based on a constitutional monarchy with representative assemblies. Although the land reform continued from 1766 to 1784, the constitutional reform never matured. Peter Leopold's reforms extended to an entire transformation of the bureaucratic and administrative state machinery, to a full-scale attack on church property and prerogatives, and to a complete judicial overhaul and promulgation of a new penal code, which was the first in Europe to abolish the death penalty. Tuscany served as a true European model of Enlightened absolutism for 25 years. But, upon the grand duke's election as Emperor Leopold II at his brother Joseph's death in 1790, Tuscany, left to his son, Ferdinand III, erupted in violence as hostile clerics and civil servants manipulated the European crisis of the 1790s against Leopoldine reform policies.

Naples and Sicily. Under Austrian Habsburg rule after 1707, Naples witnessed numerous reform plans but little concrete action. When Sicily came under Austrian rule in 1720, similar good intentions foundered in the face of local resistance, a worsening international economy, and the political exigencies and fiscal burdens of imminent wars. With the conquest of Naples in 1734 by Charles of Bourbon (ruled as Charles VII until 1759), the Italian south celebrated its independence (even if only nominally) under a new foreign dynasty. The new regime initiated a significant period of reform, which established new administrative and judicial systems, fostered economic recovery, and patronized an important Enlightenment community. The birth of the discipline of political economy dates from

The discipline of political economy

this period, with the publication in 1751 of Ferdinando Galiani's treatise *Della Moneta* ("On Money") and Antonio Genovesi's appointment in 1754 to the first university chair in political economy.

In 1759 Charles abdicated his Neapolitan throne in order to become king Charles III of Spain, leaving his minister Bernardo Tanucci to head the regency council of his son Ferdinand IV (ruled 1759–1825). The turning point in the Neapolitan reform movement came with a catastrophic famine in 1764, which urgently called into question the effectiveness of old-regime structures. After Ferdinand's marriage to Maria Carolina, the daughter of Maria Theresa, Tanucci began to lose favour with the disengaged, weak monarch. He was forced to resign in 1776, having pushed Bourbon reform to its limits, although with little tangible results. Naples moved away from its Spanish Bourbon ties and into the orbit of Habsburg policies. Fearing that the reforms had run their course, Genovesi's students in and out of government—Giuseppe Maria Galanti, Francesco Longano, Traiano Odazzi, and the Grimaldi brothers, Domenico and Francesco Antonio—pursued his interests in solving existing economic and agricultural problems. The 1780s became the high point of the Neapolitan Enlightenment, both through their work and through the writings of Genovesi's students Francesco Maria Pagano and Gaetano Filangieri. The latter's *Scienza della legislazione* (1780–85; *The Science of Legislation*), which called for equal justice for all, state intervention in economic affairs, and broad educational reforms, ranks among the most important works of the European Enlightenment. At the same time, Domenico Caracciolo, the new viceroy to Sicily from 1781 to 1785, implemented a reform program that abolished the Inquisition and questioned the fabric of the feudal system, but again without concrete results. In the end, political ties to Austria and Britain against Revolutionary France put Naples on the defensive, and, with the Neapolitan revolution, in 1799 the monarchs fled to Palermo for safety.

Gaetano Filangieri

The other Italian states. The papacy, Venice, Genoa, and Savoy eschewed political-institutional reform. The theocratic monarchy of Rome, however, was open to moderate forms of Enlightenment thought under Clement XII (reigned 1730–40) and Benedict XIV (1740–58). Under Bourbon pressure, the papacy even disbanded the Jesuits in 1773, albeit sometime after their expulsion from Portugal (1759) and from Bourbon Spain, Naples, and Parma (1767–68). Venice and Genoa lost ground as international powers and remained subject to a shrinking, conservative patriciate. Venice, however, remained Italy's most important publishing centre and home to a lively literary and artistic culture including such figures as the dramatist Carlo Goldoni and the painters Giovanni Antonio Guardi and Giovanni Battista Tiepolo. In Piedmont and Sardinia the long reign of Charles Emmanuel III (ruled 1730–73) further developed the Savoyard militaristic absolutism and centralization imperative without sharing the spirit of Enlightened reform.

The crisis of the old regime. The French Revolution did not create the European-wide crisis that followed in its wake; rather, the revolutionary repercussions that rocked polities and societies after 1789 had a long gestation period in the unaddressed problems of the old regime. Enlightenment thought, as exported from France or Austria in its Bourbon or Habsburg models, circulated freely in Italy, and a wide range of Italian intellectuals and ministers contributed to the growing body of Enlightenment thought and practice. However, this cosmopolitan world informed by secularization and science confronted powerful feudal and ecclesiastical estates who controlled vast land and wealth, combated bad government grown habitually resistant to rationalization, struggled with the difficult task of reforming a retrograde economic system unsupportive of trade or industry, and, at the same time, found itself out of touch with the daily concerns of the mass of society facing economic hardship and wedded to traditional religious beliefs. Enlightenment culture ultimately exacerbated the problem of reform, since reform from above highlighted the disparities between high and low, raised unrealizable expectations, and imposed solutions that rarely overhauled the structure of institutions. The inequalities in Italian society, the obstacles to its economic development, and the political conservatism of its privileged interest groups were not to be easily removed by reason alone.

(J.A.M.)

Revolution, restoration, and unification

THE FRENCH REVOLUTIONARY PERIOD

When French troops invaded Italy in the spring of 1796, they found fertile ground for the revolutionary ideas and practices of their native country. Since the 1780s, Italian newspapers and pamphlets had given full play to news from France, especially to the political struggle between the king and the Parlement of Paris. As the Revolution unfolded in France, news reports became more frequent and more dramatic. After 1791 they were further enhanced by the personal testimonies of political émigrés. Vigilant censorship by the Italian governments could not stop the spread of revolutionary ideas. Yet Italians had a simplistic view of the political struggle in France, one that pitted monarchists against revolutionaries.

The early years. Interest in the French Revolution was enhanced by the waning of the reformist impulse of the 1780s in the Italian states. Educated landowners and entrepreneurs who had put their trust in the enlightened rulers of their own states and had looked forward to important administrative and political reforms were disappointed. The French example gave them new hope. During the 1780s, Masonic lodges had begun to replace the scholarly academies and agrarian societies as loci of discussion of constitutional issues. In the 1790s more radical secret societies emerged, modeled after the Illuminati ("Enlightened Ones") founded in Bavaria by Adam Weishaupt, a professor of canon law, which promoted free thought and democratic political theories.

The Italian governments opposed French revolutionary ideas, recognizing them as a potential threat to stability. The kingdom of Sardinia-Piedmont therefore joined the First Coalition, an alliance formed in 1793 by powers opposed to Revolutionary France, after French armies had occupied Savoy and Nice in 1792. The arrival of the French fleet in the Bay of Naples in December 1792 prevented the king of Naples from following the Piedmontese example, but other governments resorted to stern repression of French-inspired protests. Many Italians, however, viewed the French legal and administrative system as the only answer to their own grievances against traditional elites. In Piedmont and Naples, where discontent was especially widespread, actual conspiracies were organized by proponents of democratic ideas. Arrested conspirators who escaped death or jail found refuge in France, where they became influential and active.

Italian émigrés helped to give a sharper focus to the aims of revolutionary protest and to prepare the ground for French intervention in the peninsula. The best-known émigré, the Tuscan nobleman Filippo Buonarroti, served as national commissioner in the Ligurian town of Oneglia, captured by French armies in 1794. Oneglia became the location for the first revolutionary experiment on Italian soil, when Buonarroti introduced a republican constitution and the cult of the Supreme Being and abolished seigneurial rights. The "Oneglia experiment" ended abruptly in 1795 with the fall of Maximilien Robespierre's government in France, but Buonarroti persisted in his radical beliefs, becoming a supporter of the French left-wing agitator François-Noël (Gracchus) Babeuf. The example he had set was not forgotten.

The Oneglia experiment

French invasion of Italy. The French campaign in Italy, which assured the political future of Napoleon Bonaparte, began in March 1796. By the Peace of Paris (May 15, 1796), King Victor Amadeus III of Sardinia-Piedmont was forced to cede the transalpine provinces to France and to grant safe passage to the French armies. On the same day, Napoleon's army drove the Austrians out of Milan, pursuing them into the territory of the republic of Venice. Thus, by April 1797 the French controlled the entire Po Valley, including Bologna and the northern reaches of the Papal States, which had been ceded to them by the pope in the Peace of Tolentino (Feb. 19, 1797). French armies also occupied the duchy of Modena and most of the grand duchy of Tuscany, including the port of Livorno. After defeating the Austrians on Venetian territory during the winter of 1796–97, Napoleon turned his offensive northward, crossing the Tagliamento River and driving for the Habsburg capital, Vienna. In April 1797, at Leoben, Austrian envoys offered to negotiate. In exchange for an end to the French offensive, Austria agreed to partition Venetia and to recognize French sovereignty over Belgium and Lombardy. In the following two years the Italian peninsula enjoyed a period of relative freedom and democracy, which ended with the Austro-Russian campaign against France in April 1799.

Roots of the Risorgimento. The origins of the Italian Risorgimento—the great national "resurgence" of the 19th century—may be traced to this period. It definitely dates from this time, insofar as it involved the formation of political groups affirming the right of the Italian people to achieve a government suited to their desires and traditions as well as the growth of a sense of nationalism and individual responsibility.

Italian Jacobinism

In 1800 the Neapolitan historian Vincenzo Cuoco argued that the Italian revolution of the 1790s had been a "passive revolution," having neither real roots in Italian soil nor a national ruling elite. Later generations of historians repeated and endorsed this view, arguing that Italian Jacobinism had imitated Robespierre's ideology. They tried to distinguish between Jacobin republicanism and more moderate, indigenous political movements. However, in reality, the Italian Jacobins often modified their positions from political necessity. Some who had advocated radical republicanism and democracy in the 1780s and '90s accepted important offices in the Napoleonic governments of the early 1800s. The essential difference between moderate and radical Francophiles lay in the different meanings given by each group to the concept of popular sovereignty. The doctrine of equality, for instance, could be restricted to equality before the law, or it could be expanded to include social and economic equality, which would shake the foundation of private property. Depending on their definition of equality, the two groups could take very different approaches to taxation, regulation of economic activity, and public education.

The Italian republics of 1796–99. During the revolutionary triennium (1796–99), political initiative in Italy remained in French hands. The moderate heads of the post-Jacobin Directory regarded the conquered Italian territories primarily as bargaining chips. However, Napoleon, as commander of the French armies in Italy, worked actively to establish "sister republics." He hoped for financially stable and politically dependable governments that would recognize French hegemony and adopt French legislation while holding radical elements at bay. Thus, he supported the establishment of moderate republican governments headed by prominent Italian citizens.

The Cisalpine Republic

The first of these, the Cispadane Republic, was established at Modena in March 1797; in July it merged with the Cisalpine Republic, which encompassed most of Lombardy. Although strong enough militarily to deter an Austrian offensive, the republic remained torn internally by strife between moderates and radicals. Democratic clubs and newspapers continued to resist control from Paris. Yet the moderates, under French tutelage, gradually emerged as a new bureaucratic and political class. A third republic, the Ligurian Republic, was proclaimed on June 6, 1797. It was ruled by members of the local aristocracy, who worked hand in hand with the Directory in Paris and blocked union with the Cisalpine Republic. In Piedmont the Savoy government suppressed Jacobin uprisings until the French forced the king to leave, annexing his territories in February 1799. When Napoleon ceded Venetia to Austria by the Peace of Campo Formio (Oct. 17, 1797), Italian revolutionaries felt outraged and betrayed; their disillusionment was expressed by Ugo Foscolo in his novel *Le ultime lettere di Jacopo Ortis* (1798; "The Last Letters of Jacopo Ortis"). To keep both a hostile Pope Pius VI and democratic clubs in check, the French occupied Rome in January 1798 and proclaimed a Roman Republic on March 15. Although the democratic Constitutional Club in Rome remained strong, moderate leaders maintained control. The southern exile Vincenzo Russo described these events in his *Pensieri politici* (1798; "Political Meditations"), one of the most important examples of Italian Jacobin thought.

The situation in Italy changed in November 1797, when Napoleon departed on his ill-fated expedition to Egypt. Under pressure from England, King Ferdinand IV of Naples invaded the Papal States and attempted to restore the papal government in Rome. The French armies launched a counteroffensive. Despite popular resistance that held them up for three days, they occupied Naples on Jan. 23, 1799, and established the Parthenopean Republic. King Ferdinand took refuge in Sicily under the protection of the British fleet. Although the Parthenopean Republic controlled only some of the provinces of the former Bourbon kingdom—others remained under Bourbon rule or in the throes of anarchy—it became the most democratic of all revolutionary governments of the triennium. This could, in large part, be attributed to the French military commander, Jean-Étienne Championnet, as well as to the commissioner, Marc-Antoine Jullien. Previously a follower of Babeuf, Jullien defied the wishes of the Directory in Paris for a moderate government. The Parthenopean Republic had the enthusiastic support of a number of southern intellectuals and notables.

The Parthenopean Republic

Collapse of the republics. Early in 1799 the political situation in Italy deteriorated rapidly. With the birth of the Second Coalition against France (March 1799), Austrian and Russian troops were able to occupy the Cisalpine Republic and to reach Turin in less than two months. Thus, the French lost the entire Po Valley. In addition, the greater part of the French army was forced to withdraw from Naples. The destruction of the Parthenopean Republic was the work of bands of peasants organized by Fabrizio Cardinal Ruffo, a faithful adherent of the king. Ruffo's bands quickly disposed of the weak democratic militia. Their Armata della Santa Fede ("Army of the Holy Faith") was the most important peasant uprising in the history of modern Italy. Invoking God and king, they devastated the castles of the aristocracy and occupied communal lands that the local barons had usurped; they also killed bourgeois leaders who had set up provisional municipal governments. The reaction against the French and the indigenous Jacobins became a great antiaristocratic movement, which the Bourbon monarchy skillfully manipulated to its advantage. Naples surrendered on June 23, 1799, and soon afterward the king returned from Sicily. At the behest of the British admiral Horatio Nelson and Queen Maria Carolina, wife of Ferdinand and a sister of Marie-Antoinette of France, the king ordered the execution of more than 100 revolutionary leaders. Among them were the best southern administrators, jurists, and intellectuals.

The French, who had occupied Tuscany between March and July 1799, were driven out by a violent peasant uprising, the Viva Maria ("Long Live the Virgin Mary"). This movement developed into a march on urban centres, assaults on Jewish residents, and a hunt for real or alleged local Jacobins; it also reestablished the power of the landowning aristocracy and of the clergy. The Roman Republic fell in September 1799. The French resisted only in Genoa, while a large number of Italian Jacobins took refuge in France. Thus ended the revolutionary triennium.

The pro-French *patrioti* (patriots) had completely failed to enlist the support of the masses. From the summer of 1796 the rural districts were in ferment, but almost always in opposition to the new rulers. There were peasant marches on cities in Lombardy, the Romagna, and Tuscany. Armed bands controlled or recaptured parts of the Marche, Tuscany, and the Kingdom of Naples. In some cities, such as Verona and especially Naples, popular dislike of the French and the local Jacobins was manifest. This antirevolutionary sentiment can be explained to some extent by the influence of the clergy and the high taxes levied by the republican regimes. However, it was primarily provoked by the populace's ingrained and instinctive conservatism, which only the gradual development of a grass-roots opposition movement would prove capable of overcoming.

The Italian Jacobins, defeated in domestic political struggles, also suffered a deep loss of respect for their French ally. Money levies for military purposes degenerated into pure plunder; constitutions were not democratically drafted but dictated by the French; supporters of the democratic opposition were jailed or removed from office. Worst of all, Napoleon became blatantly autocratic in his policy of returning the king of Sardinia-Piedmont to the throne in the summer of 1796 and of ceding Venetia to the Habsburgs in 1797. Disillusionment with French policies, however, did not reconcile the Italian Jacobins with their former rulers; instead, it bolstered their nationalism. In Piedmont, for instance, a secret society, I Raggi ("The Beams of Light"), advocated a democratic, unionist, and anti-French program that would lead Italy toward unity and independence.

The French Consulate, 1799–1804. Having become master of France by his coup d'état of 18 Brumaire (Nov. 9, 1799), Napoleon renewed his Italian campaign. His armies crossed the Alps again, this time through the difficult Great Saint Bernard Pass, and reoccupied Milan on June 2, 1800. A few days later they scored a definitive victory over the Austrians at Marengo, between the Po and Bormida rivers. Defeated also on German soil, the Second Coalition quickly collapsed. With the Treaty of Lunéville (Feb. 9, 1801), the Ligurian and Cisalpine republics were reestablished. Piedmont was reannexed to France in September 1802, together with Elba and Piombino. The duchy of Parma was also annexed, although annexation became official only in 1808. Even in Tuscany, Austrian influence was ended when Louis, son of Ferdinand of Parma, was declared king of Etruria. In northern Italy, Austria retained only Venetia, while France directly or indirectly controlled the areas from the Alps to the Tuscan Maremma. In the south the papal and Bourbon governments remained in power, but their positions were weak.

The second Cisalpine Republic, established in June 1800, soon proved to be a transitional regime, since it lacked the necessary combined support of the moderates and landowners. In Paris, Napoleon's most trusted adviser on Italian affairs was the Milanese patrician Francesco Melzi d'Eril, who during the triennium had hoped to see northern Italy united in a constitutional monarchy under a Habsburg or Bourbon prince. Melzi was the most clear-sighted exponent of an older moderate ruling class that still yearned for enlightened autocracy. Napoleon also favoured the formation of a large Italian state, provided he could control it. His preference was for an Italian republic with a constitution on the French model. Central authority was to be vested in a president, with a relatively weak representative body divided among three estates—landowners, merchants and tradesmen, and intellectuals and clerics. Napoleon wanted to assume the presidency himself or to name a member of his family to the position. At Melzi's insistence, however, the new state was not simply proclaimed by the French but was created by an Italian constituent assembly meeting in Lyon, France, in January 1802. Napoleon was elected president of the new Italian Republic, though not without opposition; and Melzi became its vice president. Melzi pursued a policy of compromise and co-optation. Although a majority of prefectures and ministries were held by notables, mostly members of the aristocracy, representatives of the democratic opposition were gradually included and given important posts. During the entire Napoleonic period, the formation of an Italian army was a principal aspiration of the republican government, and enduring nationalist sentiments emerged in military ranks. Serving as administrators and political leaders, the local nobles and educated bourgeois for the first time felt an obligation to govern and defend their country together.

Francesco Melzi

The Napoleonic empire, 1804–14. *Northern and central Italy.* Soon after Napoleon's rise to emperor in 1804, the Italian Republic became a kingdom, proclaimed on March 17, 1805. As king of Italy, Napoleon appointed his stepson, Eugène de Beauharnais, viceroy and Antonio Aldini, secretary of state, forcing Melzi to step aside. Although Italian autonomy remained limited, Napoleon's victories, which constantly increased the territory of the kingdom, provided some compensation. Venetia was annexed to it by the Treaty of Pressburg (Dec. 26, 1805), and Dalmatia and Istria were joined with a separate constitution. In a reorganization following the Treaty of Schönbrunn (Oct. 14, 1809), Dalmatia and Istria were joined with Trieste and Ragusa (now Dubrovnik), together with other territories ceded by Austria, to form the seven French *départements* of the Illyrian provinces. The Marche became part of the Italian kingdom in April 1808; Liguria was directly annexed to France on June 4, 1805; and Tuscany was added to it in January 1806, although this act did not become official until March 2, 1809. In 1809 Napoleon abolished the temporal power of the papacy. Pope Pius VII responded by excommunicating Napoleon, who, in response, held the pontiff prisoner, first in France and later in the Ligurian town of Savona.

As emperor of France and king of Italy, Napoleon directly controlled all of northern and central Italy. During his rule, far-reaching reforms were instituted. Although the new Italian legal codes were translated almost verbatim from the French, with little regard for Italian traditions, they introduced a modern jurisprudence responsive to the rights of the individual citizen. Properties held in mortmain, the old feudal ecclesiastical tenure (specifically those of the regular clergy), were transferred to the state and sold. The remaining feudal rights and jurisdictions were abolished. Roads were improved everywhere, and both primary and higher education were strengthened. The increased pressure of taxation was thus compensated for by a network of new and improved services that were to hasten Italian social and economic progress and cohesion.

Napoleonic reforms

The Continental System, a blockade designed to close the entire European continent to British trade, was proclaimed on Nov. 21, 1806. It was freely violated everywhere, including along the Italian coastline. Although the blockade's real purpose was to promote the growth of French manufacturing, especially the silk industry, the war economy also stimulated Italian production, particularly the emergence of machine-building and metallurgy and the completion of important public works.

The Kingdom of Naples. In the south, after the repression and executions of 1799, the Bourbons experimented with some cautious reforms, mainly fiscal and antifeudal. These were implemented to strengthen the loyalty of the rural population, which had already proved so valuable to the monarchy. But the Neapolitan government was desperately weak, both politically and militarily. Indeed, the French were able to reoccupy the country between February and March 1806, and the Bourbon court once more fled to exile in Sicily. On March 30, 1806, Napoleon's brother, Joseph Bonaparte, was proclaimed king of Naples. When he became king of Spain, in 1808, he was replaced by one of the most famous French generals, Joachim Murat. Despite this change, the nine years of French rule in southern Italy were a period of continuity, and consequently French reforms had a lasting impact. Joachim Murat was more independent of Paris than Joseph Bonaparte had been. During his reign, there were fewer French ministers and advisers in proportion to Neapolitan officials, and he opposed the enforcement of the Continental System. Feudal privileges and immunities were finally abolished, though the landed aristocracy retained extensive power in the countryside. By purchasing the property confiscated from the church and from exiled landowners, southern notables subverted Murat's plan to distribute small landholdings to peasant families. Much common land, originally usurped by large landowners, was recovered, but this worked to the benefit of bourgeois notables known in the south as *galantuomini* (honourable men). Fiscal, judicial, and educational reforms, similar to those introduced in the Kingdom of Italy, were implemented in Naples.

Sardinia and Sicily. Meanwhile, both Sardinia, where the Savoy court took refuge, and Sicily remained apart from the Napoleonic world. In Sicily the Bourbons were under strict English control, not only military but also political. In 1811–12, disagreements between the king and the Sicilian nobles, mostly over taxation, caused the intervention of the British naval commander, Lord William Bentinck. He introduced a moderate constitution that left much power in the hands of the nobles but markedly limited the absolute powers of the throne. Sicily then experienced a short period of autonomy with intense political ferment, which ended in 1816 when the restored Bourbons abrogated the constitution and reunited the island with the Kingdom of Naples, thus restoring the Kingdom of the Two Sicilies.

The end of French rule. The Napoleonic regime collapsed in Italy as it did in the rest of Europe. Beauharnais and Murat, with their respective armies, had taken part in Napoleon's disastrous Russian campaign of 1812. At the moment of defeat, Murat deserted the emperor, returned to Naples, and made peace with the English and the Austrians. Joining them in their campaign against Beauharnais, though without a full commitment, he advanced with his Neapolitan troops as far as the Po River (March 1814). By the terms of the armistice of Schiarino-Rizzino (April 16, 1814), Beauharnais was able to retain control of Lombardy. But an insurrection in Milan on April 20 allowed the Austrians to occupy the entire region.

THE RESTORATION PERIOD

The Vienna settlement. The Congress of Vienna (1814–15), held by the victorious allies to restore the pre-Revolutionary European political status quo, determined that the Bourbons should be returned to Naples. For this reason, taking advantage of Napoleon's escape from Elba to France on March 1, 1815, and his return to power, Joachim Murat opted to change sides yet one more time and declared war on Austria on March 15, 1815. By the Rimini proclamation on March 30 he incited all Italian nationalists to war, but no general insurrection occurred. Quickly defeated, Murat was forced to abdicate in May. From his exile in Corsica he moved to a base in Calabria in order to attempt the reconquest of his kingdom. Recaptured by Bourbon troops, he was executed in October 1815.

The Vienna settlement.

The Congress of Vienna established the political order in Italy that lasted until unification in 1870. Francis I of Austria also became king of Lombardy-Venetia, which was incorporated into the Habsburg state. The former episcopal principality of Trento was formally annexed to Austria. King Victor Emmanuel I of Savoy recovered his territories (Nice, Savoy, and Piedmont) and acquired the Ligurian coast. The duchy of Parma was granted to Marie-Louise of Habsburg, the daughter of Francis I and Napoleon's second wife. At her death the duchy was to revert to the Bourbon-Parma family, which was also temporarily placed in charge of the duchy of Lucca. The Habsburg-Este family returned to Modena and inherited the duchy of Massa in 1825. Also in Tuscany, the Habsburg-Lorraine family added the State of the Garrisons to its former domains and was given claim to Lucca, which the Bourbon-Parma family was to relinquish in 1847. The pope recovered his temporal domain in central Italy. Ferdinand IV of Naples once again assumed control of his former realm under the new title of Ferdinand I, King of the Two Sicilies.

Austrian hegemony restored

Thus, the Vienna settlement had brought about the dismantlement of the three aristocratic republics of Venice, Genoa, and Lucca; it had strengthened Piedmont and restored undisputed Austrian hegemony in the peninsula. Austrian troops garrisoned Ferrara, ready to intervene in case of trouble in the Papal States. Austria was given the right to intervene in the Kingdom of the Two Sicilies, if necessary. Members of the house of Habsburg ruled over Parma, Modena, and Florence; and Venetia and Lombardy became, in practice, provinces of the Austrian Empire. Only Piedmont remained outside the Austrian system designed and imposed on Italy by the Austrian foreign minister Klemens, Count von Metternich. Under Russia's secret protection, the Savoy government proved dependably reactionary.

On April 7, 1815, Francis I proclaimed the formation of the Kingdom of Lombardy-Venetia. The new state was a fiction, however, because the two regions remained separate, each subject to the central ministries in Vienna. Milan lost its role as a capital, most of the Napoleonic administration was dismantled, and the centralizing authority of Vienna became all-pervasive. Many reforms, especially those in jurisprudence, were abolished. Discontent proved general, and Austria reacted with increasingly severe police measures and stricter censorship, suppressing, for example, the liberal and Romantic periodical, *Il conciliatore* ("The Conciliator"), after only one year of publication (1818–19).

Returning to Piedmont from exile in Sardinia, Victor Emmanuel I of Savoy abolished all laws promulgated by the French and removed from public office all those who had collaborated with them. He invited the Jesuits back into the kingdom and turned many educational institutions over to them and to other religious orders. But this extreme reaction provoked liberal opposition among enlightened members of Piedmont's upper classes.

Francis IV of Modena demonstrated comparable intransigence; but, in Parma, Marie-Louise of Habsburg practiced political moderation and preserved many French reforms. Although Francophiles were expelled from the Tuscan administration and some French reforms were abolished, Tuscany under Ferdinand III of Habsburg-Lorraine and his successor, Leopold II, became known for economic liberalism and lenient censorship. Intellectual life flourished with the arrival from other regions of exiled writers such as the poets Giacomo Leopardi and Niccolò Tommaseo and the historian Pietro Colletta. These men gathered around the Gabinetto di Lettura ("Literary Club") of Gian Pietro Vieusseux, founder of an important periodical, *L'antologia* (1821–33; "*The Anthology*").

In the Papal States the Restoration, achieved principally by the diplomacy of the cautious secretary of state, Ercole Cardinal Consalvi, was characterized by increasing government centralization. Educated men who had held positions of responsibility under the French and Italian governments resented bitterly the restoration of clerical control over all aspects of public life. Dissatisfaction was especially strong in the Romagna.

In Naples the victorious powers made sure that the Bourbons would not repeat the reprisals of 1799. Thus, the Restoration appeared to begin well under the balanced policies of a government led by Luigi de' Medici, who absorbed part of Murat's capable bureaucracy. Many judicial and administrative reforms of the French era were retained, but concessions made to the church in a concordat concluded in 1818, as well as financial retrenchment, hampered the progress of the bourgeoisie. Especially among the *galantuomini,* who had profited from French legislation, discontent was strong and eventually found an outlet in a widespread secret society, I Carbonari ("The Charcoal Burners"). Already in existence under French rule, apparently with a vaguely nationalist program, the society gained strength and formulated more definite constitutional aims. The southern bourgeoisie was determined to take part in political life and to promote its interests openly. From the south the lodges of the Carbonari quickly spread to the Marche, the Romagna, Piedmont, and Milan.

I Carbonari

Spain experienced a revolution in 1820, in which the liberals gained power and reestablished a constitution promulgated in 1812. This event had notable repercussions in Italy. In the Kingdom of the Two Sicilies, former members of Murat's army, affiliated with the Carbonari, marched on Naples (July 2, 1820) to the cry of "Long live liberty and the constitution." They found support in the army and among the bourgeoisie. King Ferdinand was forced to yield to demands for the introduction of the Spanish constitution, which limited royal powers, decreased centralization, and reduced the influence of the capital. The new regime proved short-lived, however, for it had too many enemies. The king sought to recover his former powers; and Sicilian dissidents attempted to reestablish their island's separate status, though their movement was brutally suppressed by the Neapolitan constitutional government, assisted by Austria. At the Congress of Vienna, Austria had been conceded the prerogative of intervening, if necessary, to maintain the restored Bourbon monarchy. Thus, in January 1821 Metternich convened an international congress at Laibach (Ljubljana) attended by representatives of the European powers and of the Italian states and by King Ferdinand himself. Overcoming weak Anglo-French opposition, Ferdinand obtained approval for military intervention. Accordingly, the Austrian army entered the kingdom and occupied Naples on March 23, 1821, reestablishing the king's absolute government.

In Piedmont the more liberal and educated wing of the nobility resented Victor Emmanuel I's reactionary policies and found allies among bourgeois groups that had adopted the constitutional program of the Carbonari. In the wake of the Neapolitan revolution, a conspiracy began with the support of liberals in Lombardy and, covertly, of the heir apparent to the throne of Sardinia-Piedmont, Charles Albert, Prince di Carignano. Between March 9 and 13, 1821, the revolt, organized by military and bourgeois leaders, spread from Alessandria to Turin. Victor Emmanuel I abdicated in favour of his brother, Charles Felix, and, in the latter's absence from the kingdom, appointed Charles Albert as regent. On March 14, Charles Albert proclaimed the Spanish constitution of 1812, though its implementation was contingent on the new king's approval. From his refuge in Modena, Charles Felix refused to accept it; with Austrian help and loyal Piedmontese troops, he moved quickly to reoccupy the kingdom and reestablish his authority. Three conspirators were executed and many more imprisoned or exiled. Charles Albert succeeded in reconciling with Charles Felix, but his vacillating conduct marked him for years to come. The liberals never forgave him his compromises with Charles Felix, who ruled until 1831.

Although there was no revolution in Lombardy-Venetia, a complex network of opponents of the regime was discovered and suppressed. In October 1820 the Carbonari in Milan were attacked, and some were deported. In March 1821 the police penetrated another secret organization, I Federati ("The Confederates"), led by the Milanese nobleman Federico Confalonieri. The society favoured constitutional government, but its program was more moderate

I Federati

than that of the Carbonari though no less anti-Austrian. From December 1821 to January 1823 members of the conspiracy were unmasked in the army and the upper bureaucracy and received death sentences, all of which were eventually commuted to long prison terms.

Economic slump and revival. Political reaction, which continued in the Romagna until as late as 1828, was accompanied during this decade by a severe economic recession. After the famine of 1816–17, Russian grain flooded the Italian market and contributed to a crisis of agricultural overproduction. The desperate poverty of the peasantry led to grain riots, brigandage, and the spread of pellagra, a vitamin-deficiency disease endemic among the corn-eating northern peasantry. The slump continued until nearly 1830, when successful mulberry cultivation brought renewed rural prosperity and was sufficient, particularly in Piedmont and Lombardy, to reestablish agricultural credit and provide capital for the growth of textile and engineering industries.

Renewed prosperity made possible a revival of cultural activities; the country's economic and social problems became the subject of discussion in many periodicals. The most important of these was the philosopher Gian Domenico Romagnosi's *Annali universali di statistica* ("World Statistical Almanac"), to which his most important pupil, Carlo Cattaneo, contributed his first essays. In fact, the ranks of political and cultural critics, hitherto dominated by Lombard and Tuscan moderates, were augmented by expatriates from other regions and by Roman Catholic and democratic thinkers.

The rebellions of 1831 and their aftermath. The Parisian July Revolution of 1830 set in motion an Italian conspiratorial movement in Modena and in other Emilian towns. Two Carbonari, Enrico Misley and Ciro Menotti, put their trust in the duke of Modena, Francis IV of Habsburg-Este, who was looking for an opportunity to expand his small state. But when Francis discovered that the plot was known to the Austrian police, he had Menotti and others arrested. Nevertheless, the revolt spread to the Romagna and to all parts of the Papal States except Lazio. For various reasons, the provisional governments of the insurgent cities failed to organize for a common military defense and did not receive the hoped-for help of the French army. During March 1831 the Austrian army intervened and reestablished the status quo ante. The failure of the uprisings of 1831 suggests that the program of the Carbonari had run its course.

The moderate liberals, most of them Carbonari, had demonstrated a readiness to compromise with the absolute monarchs. They had distrusted democrats and republicans who sought to achieve Italian unification by political revolution and force of arms. Among these were the Adelfi, a secret society of the followers of Filippo Buonarroti. Ultimately, the task of organizing new cadres of democratic and republican opponents of the Restoration governments fell to Giuseppe Mazzini, scion of a bourgeois and Jacobin family of Genoa. Exiled in 1830 at the age of 25, Mazzini turned away from both *Carboneria* and Buonarrotism and established his own organization, La Giovine Italia (Young Italy). Republican and unionist, Mazzini's organization emphasized popular participation in the national struggle but eschewed Jacobin and social-revolutionary objectives. In 1833–34 the first abortive Mazzinian uprisings took place in Savoy and Genoa. The latter was organized by Giuseppe Garibaldi, who then fled to France. In 1834 the Austrian police identified as many as 2,000 adherents of Young Italy in Lombardy. In 1836 Mazzini, who had established relationships with democratic revolutionaries in other countries and cofounded La Giovine Europa (Young Europe), left Switzerland and settled in London.

Giuseppe Mazzini

Conservative repression convinced the moderates that to organize conspiracies with limited membership was futile and that instead it was imperative to educate the public about the need for change. Meanwhile, the peace imposed on the Italian peninsula from 1831 to 1848 favoured economic development, which had come in varying degrees everywhere except in the south. There the Bourbon Kingdom of the Two Sicilies remained backward, and the growth of bourgeois landownership that resulted from the division of great feudal holdings did nothing to change the situation. Thus, the imbalance between north and south, to be felt even more strongly after unification, continued to increase. Meanwhile, Genoa, Turin, and Milan began to lay the foundation for becoming important European financial and industrial centres. Piedmontese and Lombard manufacturing as well as banking expanded rapidly. In Venetia important land-reclamation projects were completed; and in Tuscany banking and commerce flourished, especially via the port of Livorno. Throughout the country the construction of a railroad network increased commerce and gave rise to subsidiary industries.

Economic revival made it more difficult for governments to tighten police control. In Milan, Carlo Cattaneo's journal, *Il politecnico* ("The Polytechnic"), founded in 1839, argued that the progress of science and technology necessary to fuel economic growth depended upon government reforms. In the same year, a congress of Italian scientists held its first annual meeting in Pisa. Through 1847 each subsequent meeting assumed a markedly more nationalistic character. Thus, conditions became more favourable for moderates to realize their programs of increasing public education and abolishing censorship and police surveillance. In the cause of economic unification they endeavoured to standardize tolls and trade practices and to increase cultural exchange among the Italian states. They also sought to achieve representative institutions compatible with Italian traditions and with Roman Catholicism.

Vincenzo Gioberti

Vincenzo Gioberti, the most important exponent of liberal Catholicism, envisioned a new and positive role for the temporal power of the papacy. His *Del primato morale e civile degli Italiani* (1843; "On the Moral and Civil Primacy of Italians") affirmed the idea of progress as the return of the material to the spiritual, of man to God. Because such progress could be realized only through the mediation of the church, Gioberti advocated an Italian federation free from Austrian hegemony and under the nominal presidency of the pope. His ideas were influential among the clergy and most Catholic intellectuals. Under different formulations, this new papalist movement was advanced by Cesare Balbo, Niccolò Tommaseo, and Antonio Rosmini-Serbati.

In the early 1840s, renewed Mazzinian attempts at armed rebellion were ruthlessly suppressed. Among these was the Calabrian expedition of 1844 organized by the Venetian Bandiera brothers and seven of their companions, who were captured and executed by the Bourbon regime. These violent acts of suppression increased the esteem felt by governments and the general public for the moderate opposition. The election of Giovanni Maria Mastai-Ferretti as Pope Pius IX in 1846 augured well for the Papal States, his nomination being the result of anti-Austrian feeling in the Curia. In the beginning of his reign, he showed liberal sympathies and granted amnesty to political prisoners. He gradually removed the most reactionary prelates from important government posts, permitted the publication of political periodicals, and finally, in 1847, established a council of state. Although only advisory, the council gave the laity a voice in the affairs of state. Influenced by the pope's liberalism, rulers elsewhere in Italy introduced reforms. Especially important was the Tuscan press law of 1847, by which Grand Duke Leopold II removed most forms of political censorship. The reforms encouraged extremism, however, and the reactionary powers of Europe became convinced that the stability of Italy was in jeopardy. In July 1847 Austrian troops occupied the papal city of Ferrara. This intervention stimulated cooperation among Italian rulers, including Charles Albert of Savoy, whose relations with Austria had been particularly strained. While the rulers discussed reforms—especially the formation of an all-Italian customs union—and the measures needed to cope with famine in several regions, the populace began to stir.

The revolutions of 1848. The first of the revolutions of 1848 erupted in Palermo on January 9. Starting as a popular insurrection, it soon took on overtones of Sicilian separatism and spread throughout the island. Piecemeal reforms proved inadequate to satisfy the revolutionaries, both noble and bourgeois, who were determined to have a new and more liberal constitution. Ferdinand II of the Kingdom of the Two Sicilies was the first to grant one (Jan. 29, 1848). Other rulers were compelled to follow his example: Leopold II on February 17, Charles Albert on March 4, and Pope Pius IX on March 14. The Austrian government, on the other hand, did not yield to popular pressure. Instead, it reinforced its garrisons in Lombardy-Venetia, arrested opposition leaders in Venice and Milan, and suppressed student demonstrations in the university cities of Padua and Pavia. By March 22–23, when revolution had also reached Budapest and Vienna, Venetian and Milanese insurgents moved to depose their Austrian overlords. Within a few days the Austrian army lost nearly all of Lombardy-Venetia and retreated into the Quadrilateral (the region between Mantua, Verona, Peschiera, and Legnano).

War between Sardinia-Piedmont and Austria

On March 23 Charles Albert of Sardinia-Piedmont declared war on Austria. It was a risky decision, but prospects for a national war seemed promising; and he wanted to seize the initiative to preclude republican and democratic domination of the insurgency. Annexing Parma and Modena, whose rulers had been driven out by insurgents, the Piedmontese won a few more victories before suffering reverses. Pius IX, Leopold II, and Ferdinand II, all of whom had initially sent troops to northern Italy to support the Piedmontese army, hastily withdrew their forces. The pope's address to the cardinals on April 29 revealed his reluctance to back national movements against Austria and did much to discredit him among patriots. Lombardy and Venetia, though not without internal opposition, accepted merger with Piedmont. Nevertheless, the Piedmontese army was unable to withstand the Austrian counteroffensive. After a series of defeats, Charles Albert's forces withdrew from Milan. On August 6 they crossed the Ticino River, leaving the city and its insurgents to the mercy of the returning Austrians. Accusations of royal treachery, formulated by Lombard democrats at that moment, long survived in Italian political debates. By the terms of the Salasco armistice (Aug. 9, 1848), the Piedmontese army abandoned Lombardy. In Piedmont the new constitution, the Statuto (Statute), remained in force, and democratic ideas survived.

Throughout Europe, the forces of reaction were triumphant. The revolutions of 1848 were suppressed in Vienna, Prague, Budapest, and Paris. In Naples the king regained power in a coup on May 15 and subsequently went on to reconquer Sicily. Meanwhile, in Rome a range of obscurantist policies were reintroduced. Venice, however, under the dictatorship of Daniele Manin, refused to accept the Salasco armistice and resisted the Austrian siege. Leopold II of Tuscany took refuge in the Bourbon fortress of Gaeta in February 1849, when the democrats Giuseppe Montanelli and Francesco Domenico Guerrazzi were on the verge of taking control of the government and proclaiming an Italian constituent assembly. In Rome the minister Pellegrino Rossi, a former member of the Carbonari who had promoted conciliatory policies after returning from exile in France, was assassinated on Nov. 15, 1848. This event triggered a democratic insurgency and caused Pius IX to flee to the safety of Gaeta. A constituent assembly elected by universal male suffrage proclaimed the Roman Republic on Feb. 5, 1849.

Piedmontese defeat

The Italian revolution seemed to have been reborn. However, Charles Albert, pressed by Piedmontese democrats to resume his war with Austria (March 20, 1849), saw his army routed at Novara three days later. On the same day, March 23, he abdicated and went into exile. His successor, Victor Emmanuel II, was granted an honourable armistice because the Austrians did not want a weakened Savoy monarchy that could be exploited to the advantage of its democratic opponents. The defeat of Piedmont made the position of the democratic and republican opposition untenable in other parts of Italy as well. In Tuscany moderates recalled the grand duke, whose Austrian protectors crushed an insurrection in radical Livorno (May 1849). In Lombardy the Austrian reconquest of Brescia in March, after 10 days of fighting, left Venice isolated, though the city resisted enemy forces until August. The Roman Republic, led by Mazzini and Garibaldi, held out until July 3 against a French army sent by Louis-Napoléon Bonaparte, the new president of the French Republic, whose restoration of the papacy thus repaid his Roman Catholic supporters. The returning sovereigns rapidly set about abrogating constitutions, disbanding parliaments, and, especially in the south, filling the prisons.

UNIFICATION

The role of Piedmont. In Piedmont Victor Emmanuel II governed with a parliament whose democratic majority refused to ratify the peace treaty with Austria. This was an exception to the general course of reaction. The skillfully worded Proclamation of Moncalieri (Nov. 20, 1849) favourably contrasted Victor Emmanuel's policies with those of other Italian rulers and permitted elections. The victorious liberals installed a new cabinet under Massimo d'Azeglio, a moderate trusted by the king. D'Azeglio introduced the Siccardi law, which curtailed the power of ecclesiastical courts. In October 1850 another prominent moderate, Camillo Benso di Cavour, entered the cabinet and from that time forward directed a laissez-faire economic policy. He formulated international commercial treaties and drew on foreign capital to reduce the public debt, stimulate economic growth, and develop a railroad system. Cavour's dynamism was viewed with suspicion by conservatives and even by d'Azeglio. In 1852, through an alliance with centre-left deputies that became known as the *connubio* ("marriage"), Cavour displaced d'Azeglio as head of the cabinet. Despite disagreements between Cavour and the king (who was influenced by the clerical party and occasionally displayed absolutist tendencies), various ecclesiastical, judicial, and fiscal reforms were introduced.

Piedmont's prestige in Italy and abroad was reinforced by a number of events. In March 1854 France and England intervened in support of the Ottoman Empire against Russia in the Crimea. To obtain Austrian support, they were prepared to guarantee the status quo in Italy. Only Piedmont was in a position to disrupt it at that time, and Cavour negotiated an alliance with the Western powers. In May he sent to the Crimea an army that performed brilliantly. As a result, Piedmont was able to assume a place among the victors at the Congress of Paris (February 1856). From this platform Cavour, achieving a diplomatic coup for Piedmont and Italy, declared that the only threat to peace in Italy, and the root cause of subversive plots, was the burdensome Austrian overlordship.

The Sapri expedition

Meanwhile, Mazzini's democratic and republican movement was crumbling. In February 1853 an insurrection against the Austrians failed in Milan. The discovery and execution at Belfiore (1852–53) of the leaders of a conspiracy in Mantua, as well as abortive insurrections in Cadore and Lunigiana, discredited the democratic movement and discouraged its most dedicated adherents. Mazzini's isolation was complete when it became known that he had supported an expedition to the southern mainland to incite insurrection, known as the Sapri expedition (June–July 1857), in which the Neapolitan republican and socialist Carlo Pisacane and some 300 companions lost their lives. The democrats were divided and unable to carry on the revolutionary struggle; nothing was to be expected from the restored governments. In Lombardy-Venetia, Austria carried out stern repressive measures. Pius IX, now under the influence of the reactionary Giacomo Cardinal Antonelli, refused to grant any reforms in Rome. Liberal Catholicism could not remain viable without reforms in the Papal States. In Naples and the duchies, reaction became pervasive, though the grand duke of Tuscany sought to make his subjects forget that he owed his throne to Austrian military intervention. Only in Piedmont was there any hope left for the reformers.

The war of 1859. In 1857 Cavour sponsored the formation of the monarchist-unionist Italian National Society. Under the presidency of Manin and the vice presidency of Garibaldi, the society achieved wider appeal than it would have achieved had it been led entirely by moderates. Although not outlawing conspiratorial movements, Cavour was determined to solve the Italian question by international politics rather than by revolution. At a secret conference held at Plombières, Fr., in July 1858 he arranged with Louis-Napoléon Bonaparte (now Emperor Napoleon III) for French military intervention in the event of Austrian aggression against Piedmont. Cavour's goal was the complete expulsion of Austrian troops from the peninsula. The price for this help was the cession of Savoy and Nice to France and the outlawing of the Mazzinian movement; wrongly, Napoleon III held Mazzini's followers responsible for an attempt on his life made by the anarchist Felice Orsini in Paris on Jan. 14, 1859. Despite that event, a Franco-Piedmontese alliance was sealed in January 1859. With Napoleon's approval, Victor Emmanuel II made a speech from his throne in which he declared himself ready to hear "il grido di dolore" ("the cry of woe") against Austrian oppression that arose from every part of Italy.

Meanwhile, the Austrian military leadership and its sympathizers at court urged Francis Joseph to declare war on Piedmont. On April 23 an insulting and unacceptable ultimatum demanded the demobilization of Piedmontese troops. Piedmont rejected the ultimatum, and Austria declared war three days later. As Cavour had hoped and planned, France honoured its alliance with Piedmont. In June 1859 the allies won bloody battles at Magenta, Solferino, and San Martino. But, with the Austrian army in retreat, Napoleon III suddenly signed an armistice with the Austrians at Villafranca. This sudden change of policy was prompted partly by the outcry of French public opinion against the loss of life in the Italian campaign and partly by events in Italy itself, where political unification seemed imminent. On April 27 Leopold II of Tuscany had been overthrown by insurgents, and moderate political leaders headed by Baron Bettino Ricasoli had formed a provisional government. In June Parma, Modena, and the Papal Legations (the northern Papal States) had rebelled. Only in the Marche and Umbria were papal troops able to suppress the insurgents. Plebiscites in the liberated states urged unification with Piedmont, but France opposed this outcome.

The Armistice of Villafranca

At Villafranca Napoleon III agreed to accept the cession

The unification of Italy. The dates are those of annexation, first to Sardinia-Piedmont and after 1861 to the Kingdom of Italy.

of Lombardy from Austria, which he passed, in turn, to Piedmont. He also agreed that the legitimate rulers of Modena and Tuscany would be restored to power and permitted to join an Italian confederation. These events marked a serious political defeat for Cavour, who resigned in July 1859 and was replaced by Urbano Rattazzi. England, however, was opposed to the restoration of conservative governments in Modena and Tuscany, and Napoleon III, with his position at home strengthened by the acquisition of Savoy and Nice, reconsidered his position. As a result, Cavour's policy prevailed, and he was returned to office on Jan. 21, 1860. New plebiscites in the duchies and the Papal Legations reconfirmed popular sentiment in favour of union with Piedmont. It was fear of a democratic revolution, the need to weaken Austria, and England's desire to set up a strong Italian state as a counterweight to French influence that induced the European powers to assist the Piedmontese monarchy in obtaining this great success.

Garibaldi and the Thousand. The democratic movement refused to acknowledge that the national revolution was in any way complete so long as parts of the peninsula remained under the old sovereigns. Sicily, where autonomist opposition to the Bourbon government was endemic and extreme, was the most obvious place for a democratic revival. In April 1860 a Mazzinian-inspired insurrection broke out in Palermo (the Gancia revolt), and, although it was quickly quelled, it spread throughout the island. After the insurrection, Sicilian democrats gave evidence that they could overcome their deep divisions of ideology and class. In May they had the opportunity to assist Garibaldi's Expedition of the Thousand, a volunteer force that had set sail from Liguria to free the Italian south from Bourbon rule.

Garibaldi's Expedition of the Thousand

Despite scant preparations and a shortage of weapons, Garibaldi's volunteers landed at Marsala on May 11, 1860, and in less than three months conquered the entire island of Sicily. Garibaldi's daring and skill and the indigenous revolutionary ferment accounted for the success of the expedition. Still, the attitude of the Sicilian peasants was ambivalent. They initially welcomed the invading force but then quickly became disillusioned because of Garibaldi's reluctance to order the breakup of secular, landed estates. Although Garibaldi declared on May 14 that he ruled "in the name of Victor Emmanuel, King of Italy," he entrusted the Sicilian provisional government to his associate, Francesco Crispi, who came into serious conflict with Cavour's emissaries on the island. Cavour feared the implications of a republican coup d'état. Meanwhile, as the European powers attempted mediation, the new king of the Two Sicilies, Francis II, granted a constitution and promised amnesty to Sicilian rebels. At this point, without the consent of Victor Emmanuel II and perhaps even against his wishes, Garibaldi crossed the Straits of Messina on Aug. 19, 1860, and by September 7 made a triumphant entry into Naples. Francis II fled to Gaeta, and on October 1 the last serious resistance of the Bourbon army was overcome at the battle of the Volturno near Caserta.

The prestige of Garibaldi and the democrats had risen so high that Cavour felt compelled to seize the initiative once again. Having persuaded Napoleon III to limit his opposition to lodging a perfunctory protest, Cavour proceeded to occupy Umbria and the Marche. Rome and its surrounding region (Latium) were to remain under the pope, but the remainder of the Italian peninsula was to become one kingdom under a moderate constitution. On Oct. 26, 1860, Victor Emmanuel II met with Garibaldi on Neapolitan soil and was greeted as "king of Italy." During October and November, plebiscites in the former papal and Bourbon provinces overwhelmingly endorsed annexation to the Kingdom of Italy.

The Kingdom of Italy

The Kingdom of Italy was officially proclaimed on March 17, 1861, by a parliament assembled in Turin. Soon afterward Cavour asserted that Rome must become the capital of the new state within a context of separation between secular and religious authority. However, because of Cavour's untimely death on June 6, 1861, the Roman question remained unresolved.

Condition of the Italian kingdom. In 1861 the kingdom had 26 million inhabitants, 78 percent of whom were illiterate. With some 70 percent of the adult population engaged in agriculture, it seemed unlikely that Italy could achieve the economic progress made by other European countries in that period. The group winning the majority in Parliament in 1861 was from the moderate-conservative right. The coalition was based principally on an alliance of a Piedmontese group, which was led by Giovanni Lanza and Quintino Sella and controlled manufacturing and banks, with Ricasoli's Tuscan group, which was interested in commerce and transportation. This elite political grouping wanted a centralized government structure that would allow Parliament, and hence the executive branch, to control local administration, especially where there was any danger of democratic predominance. By a series of laws enacted in 1865, these moderate notables effected legislative unification and established firm central control over the provinces and the communes through the appointment of handpicked regional prefects. Their democratic opposition, preoccupied with the issue of bringing Rome and Venetia into the new state, put up little resistance to these centralizing and authoritarian tendencies.

The serious economic imbalance between north and south was ultimately intensified by the kingdom's centralized administration. The free-trade policy of the moderate governments ruined or badly harmed the weak and unprotected industries of the south, especially the woolen-cloth manufacturing in the Salerno area, which had previously been sheltered under the Bourbon government. Moreover, railroad construction in southern regions was beset with corruption in the issuance of contracts. Relief for the poor and public education remained miserably inadequate. Naples, the largest city in Italy in 1861, with about 447,000 inhabitants to Turin's 205,000, was plagued with poverty and distrusted in the small provincial capitals.

Poverty was most acute and widespread in rural areas, where peasant families had gained nothing from the breakup of large feudal estates. Many peasants espoused an especially violent form of brigandage, which, though fomented and often assisted by emissaries of the exiled Francis II, was a form of class warfare against the agrarian bourgeoisie. The Italian army's savage suppression of this lawlessness resulted in the execution of at least 5,000 peasants. Even so, the brigandage was not brought under control until 1865.

In the country at large, public opinion remained preoccupied with the issue of completing territorial unification. The democrats, who were most interested in solving the Roman question, saw an opportunity to act in March 1862 when Urbano Rattazzi replaced Bettino Ricasoli as prime minister. Garibaldi assembled a volunteer force in July and August and began a march from the south into Latium and Rome. Fearful of Austrian reaction and of a military confrontation with the French garrison in Rome, Rattazzi mobilized the regular army. Garibaldi was wounded in the ensuing fighting and was arrested on Aug. 29, 1862, at Aspromonte in Calabria. The subsequent public outrage brought down Rattazzi's government. In 1864 Marco Minghetti, another moderate, negotiated a compromise solution in the September Convention, which called for the withdrawal of French troops from Rome in exchange for an Italian pledge to respect the pope's temporal sovereignty and to remain out of Rome. A secret clause in the agreement also bound Italy to transfer its capital from Turin to Florence, thus implicitly abandoning the Cavourian claim to Rome. When this clause became known, severe riots occurred in Turin (September 21–22), leaving 30 people dead and causing the fall of the Minghetti government.

The September Convention

The acquisition of Venetia and Rome. Two years later, in June 1866, the outbreak of war between Austria and Prussia diverted attention from Rome to Venice. The Italian government under Alfonso La Marmora, taking advantage of Vienna's problems elsewhere, attacked Austrian-held Venetia, only to be defeated both on land at Custozza (June 24) and at sea near Lissa (July 20). A few Garibaldian volunteers operating in the Trento area were the sole Italians to achieve some moderate success. Yet, through the mediation of Napoleon III, Italy obtained Venetia in the Treaty of Vienna (Oct. 3, 1866). In the spring of 1867, Rattazzi returned to power and gave tacit consent to the stationing of Garibaldian volunteers along the papal border. However, a renewed attempt to march on Rome merely brought back French troops, who defeated Garibaldi at Mentana on November 3. Arrested once again, he was sentenced to house arrest on the remote island of Caprera, between Sardinia and Corsica, where he owned some property. Italy suffered a marked loss of prestige politically and militarily, and the internal situation was far from favourable. Separatist revolts occurred in Palermo in 1866. In 1869 Parma and other cities rose in rebellion against the *macinato* ("grist tax") and other taxes levied to finance the reorganization of the armed forces.

The Lanza-Sella government, formed in December 1869, was perhaps the most typical among the centre-right cabinets of this period. It repressed Mazzinian opposition, advocated free trade, and was cautious in foreign affairs, carefully maintaining a pro-French bias. Yet, despite its lack of brilliance and its subservience to France (revealed in its near acquiescence to the king's desire to intervene in the Franco-Prussian war), the Lanza-Sella government resolved the Roman question. Napoleon III's defeat and abdication deprived the pope of French military protection. Therefore, on Sept. 20, 1870, following a symbolic armed resistance by the papal army, Italian troops opened a breach in the city's walls at Porta Pia and entered Rome. Refusing to accept Italy's occupation of the city, Pius IX withdrew to his palace and declared himself a prisoner in the Vatican, a position maintained by his successors until 1929. (Ma.B./C.M.Lo.)

Italy from 1870 to 1945

DEVELOPMENTS FROM 1870 TO 1914

Politics and the political system, 1870–87. After the conquest of Rome in 1870, Italian politicians settled down to manage the economy, to build up the country's military power, and—in the telling phrase of the Piedmontese author and statesman Massimo d'Azeglio—to "make Italians." Popular disaffection remained high, especially because of the grist tax that had been introduced in 1869. Governments of the "right" remained in office, first under Giovanni Lanza (to 1873) and then under Marco Minghetti (1873–76). The "right" was not an organized party but a group of patriotic, mostly northern, landowners committed to a strong currency and free trade. Under both prime ministers the main domestic task was to balance the budget. Minghetti eventually managed this, but raising taxes and squeezing expenditure had made the "right" unpopular, and its candidates did badly in the 1874 elections. In March 1876 the Minghetti government fell when its Tuscan supporters refused to support a state takeover of the railways.

The end of government by the "right"

Italy was then ruled for many years by governments of the "left," which were usually led by Agostino Depretis (until his death in 1887). The deputies of the "left," heirs of the Risorgimento's democratic tradition, were more anticlerical, more frequently members of the middle class (many of them were lawyers), more often from the south, and less concerned about the value of money than the rentier "right" had been. They were, however, splintered into various groups, and factional disputes soon became endemic. "Left" governments abolished the grist tax (1883) and made two years' primary education compulsory (1877).

A main achievement of the "left" government was the widening of the suffrage in 1882. The voting age was reduced to 21 (from 25); the requirement to pay 40 lire in direct taxes per annum was halved and abolished altogether for those with two years' schooling. The electorate thus increased from approximately 500,000 to 2,000,000 men, including now many urban artisans, especially in the north where schools were more common. Within a few years modern political parties were founded and won seats in northern Italy, but southern constituencies remained dominated by elite groups of lawyers and local notables, often linked to prominent landowners.

Local government was also very significant, and there were often bitter disputes among local factions. The 8,300-odd local municipalities (*comuni*) were in charge of primary schools and most welfare services, raised much of their own revenue, and appointed their own staff. The central government tried to control them by appointing the mayors and also by giving veto powers over municipal decisions to provincial bodies that were strongly influenced by the provincial prefect, a government official. The prefect frequently dissolved councils for alleged financial or legal abuses and replaced them with a government "commissioner" until new elections were held. This power was often used when local council leaders opposed government candidates at parliamentary elections. However, government attempts to control local government were never really successful. The prefects had to make sure government candidates would win the next parliamentary elections, and so they had to conciliate, not bully, local elites, including the mayors and municipal councillors. Corruption was therefore often left unchecked. National governments became remarkably dependent on local power-holders. Depretis himself won over ("transformed") deputies and kept his governments in office by distributing patronage and favours to local notables. *Trasformismo* soon became the normal way of conducting parliamentary business, for there were few serious disputes between the leading politicians. The constitutional settlement of 1861 was accepted by virtually all of them; foreign and colonial policy was not contentious and, in any case, was conducted by foreign ministers and prime ministers without much reference to Parliament. In 1881 the government was greatly annoyed by the French occupation of Tunisia, and in the following year, in order to avoid diplomatic isolation, Italy joined the Triple Alliance with Germany and Austria-Hungary. This was essentially a defensive alliance guaranteeing German and Austrian support against any attack by France, Italy's main rival in the Mediterranean. However, it encouraged Italy's first real colonial venture, the takeover of the Red Sea port of Massawa (Mitsiwa) in 1885. Southern politicians favoured colonial expansion as an outlet for surplus population and agricultural produce; northern ones wanted Italy to be a great power, saw the army as an essential guarantor of public order, and supported high military spending—the army and navy ministries spent more than all other ministries combined between 1862 and 1913.

Joining the Triple Alliance

Forces of opposition. The political elite may have agreed on most issues, but there was plenty of opposition in the country. Most men owned guns, and violent crime was common: there were 3,000 murders a year, many of them a result of vendettas or blood feuds. Brigands were still active in parts of the southern mainland in the 1870s, and banditry was still common in the mountainous zones of Sardinia. In the towns, rioting was frequent; more than 250 people were killed in riots against the grist tax in 1869, and similar riots against local taxes or for land and jobs continued well into the 20th century. The strikes of the 1880s—especially by agricultural labourers in Mantua province—much alarmed respectable opinion. Anarchists were active in the Romagna and parts of the south and occasionally attempted to carry out insurrections, as at Matese in 1877, or to kill the king, as Giovanni Passanante attempted to do in 1878.

However, the anarchist leader in the Romagna, Andrea Costa, soon converted to socialist ideas. In 1881 he founded the Revolutionary Socialist Party of Romagna (later the Italian Revolutionary Socialist Party), which preached eventual revolution but also agitated for such issues as universal suffrage and labour and welfare legislation; in 1882, on the new suffrage, Costa became Italy's first socialist deputy. In Lombardy a moderate, labour-oriented Italian Workers' Party was founded in 1885, which helped to organize the Po Valley peasantry into "leagues" and labour cooperatives. The northern labour movement—unions, mutual aid societies, and cooperatives—became infused with either revolutionary or reformist socialist ideas.

Republican opposition also survived, particularly in central Italy, long after Mazzini's death in 1872. Republicans ran many of the mutual aid societies and cooperatives. They opposed strikes, nationalizations, and the class struggle but strongly favoured social protective legislation and civil rights. Some of them, including Matteo Renato Imbriani, also advocated an active "irredentist" foreign policy—that is, a policy that aimed to liberate Italians living in Habsburg territory; in particular they wanted to wrest Trento and Trieste from Austrian control. They regarded the Triple Alliance and colonial expansionism as inimical to Italian interests and as expressions of Italy's monarchical and conservative political institutions.

Perhaps the most serious opposition force in the country was the Roman Catholic church. The Risorgimento had deprived the church of the Papal States, including Rome itself, and of much of its income. Its previous virtual monopoly of education and welfare had been overthrown, and compulsory state education was deliberately secular. Many religious orders had been disbanded; monasteries and convents had become public buildings, used by the state. In the south, particularly, ecclesiastical organization had relied heavily on monks and friars and could barely continue to function. Bishops were not allowed to receive their revenues and take up their posts without royal approval, which was often refused. The pope himself was permitted, by the state's "Law of Guarantees" of 1871, to retain only the Vatican and Lateran palaces as well as Castelgandolfo. Pius IX denounced the new usurping state, forbade Catholics to vote at parliamentary elections or to become candidates, and appointed a new generation of "intransigent" bishops. New laymen's organizations were founded; the Opera dei Congressi, with committees at parish level, became the focus of Catholic resistance to the new state. It organized cooperatives, welfare insurance, credit banks and mutual aid societies, as well as a host of local journals and campaigns against Liberal "lay" proposals (such as a divorce law). Church and state remained mutually suspicious, particularly in the Veneto region, where "regionalist" opposition to centralizing government and peasant hostility to landlords and free trade were both mobilized effectively by the Catholic social movement.

The "Law of Guarantees"

Land reform. The main issue of political debate in late 19th-century Italy was land. Liberal governments insisted that the municipalities should sell off most of the common land to private owners—at least 740,000 acres (about 300,000 hectares) were sold by 1880 in southern Italy alone, and more was occupied illegally. Another 1,250,000 acres of ecclesiastical estates were similarly sold, often at extremely low prices. Overall, at least 5,000,000 acres were transferred. In some regions, including Piedmont, Liguria, and Sardinia, the sales did set up a "property-owning democracy"—that is, a large number of rural people became small landowners, albeit with scattered strips that made improvement unprofitable. The sales also introduced people to the market economy, as they had to repay their mortgages in cash as well as find money for high land taxes. Small ownership did not become widespread in most other regions, despite the land sales; peasants who did acquire land were often forced to sell it again to meet tax debts or interest payments. However, land transfers did often create a nonnoble rural middle class that owned an adequate amount of land or extensive flocks and could dominate local politics; this was particularly true in the former Papal States of central Italy.

Privatization of the commons

Privatization of the commons also had serious environmental and social consequences. Much common land was woodland, bought up and felled by speculators who could sell timber to railway companies (for sleepers) or to mines (for roof support). Deforestation became widespread: Sardinia, for example, lost four-fifths of its trees in the 19th century. The results included soil erosion, landslides, stagnant water in valley bottoms, and increased malaria—the greatest scourge of rural Italy, which in turn prevented much fertile low-lying land from being cultivated. Furthermore, the state also abolished traditional rights such as grazing and wood-gathering on the remaining unsold common land. Millions of households that had relied on access to this land to provide fuel for heating and cooking or pasture for their pigs were suddenly forced either to suffer real poverty or to break the law.

Protectionism. Most agricultural land in Italy was used to grow grain, especially wheat. In the early 1880s world wheat prices fell by one-third, and the incomes of the larger and more prosperous farmers (who grew for the market rather than for their own consumption) collapsed. As landowners were the most powerful pressure group in the country and were strongly represented in Parliament, their demands for protectionism became irresistible.

Tariff protection was also favoured by the most prominent wool and cotton manufacturers of northern Italy, and these industries were second only to the silk industry in importance and numbers employed. Some tariff protection (up to 40 percent) had, in fact, already been given to textiles and other light industries in 1878, but employers naturally wanted more, particularly after the restoration of gold convertibility in 1883 in effect revalued the lira. Moreover, in the 1880s Italy also acquired a steel industry (Terni Steelworks, founded 1886), which was designed to build warships and railways but sold to subsidized industries and was itself unable to survive without protection. All this meant the rise of a strong protectionist lobby, based on large landowners and textile manufacturers and linked to powerful steel and naval interests.

Social changes. In 1871 there were 26.8 million Italians. Both birth and death rates were high, and almost half the children born alive died before the age of five. As time went on, both rates gradually declined—the first because of later marriages, the second because of improved housing, sanitation, and diet. Most people worked in agriculture but often lived in hill towns to avoid malaria or bandits; however, living in the countryside began to be more common later in the 19th century. Large-scale transatlantic emigration began in the 1880s; in 1888 more than 200,000 Italians went to the Americas in search of jobs, 10 times as many as a decade previously. Most emigrants, whether bound for the Americas or for other parts of Europe, were northerners, often seasonal migrants from hill areas of peasant ownership, where jobs were scarce and where younger sons who stayed behind had little prospect of marriage. But even in 1888 more than a quarter of the emigrants were southerners, and the great exodus of southern emigrants to both North and South America was just about to begin. Most people (68.8 percent in 1871) were illiterate and usually spoke only dialect. Illiteracy was particularly common among women, peasants, and southerners. The compulsory schooling law of 1877 was widely ignored in practice; in any case, it provided for only two years of schooling, not enough to guarantee the ability to read and count. Conscripts were likely to be taught to read during military service, but only one-quarter of the age group was actually called up into the army. Italian education was more successful at the secondary level in the towns: the "technical schools" and "technical institutes" taught science, engineering, and accounting and had high prestige among urban parents. As for the universities, they essentially trained lawyers and doctors, both professions in which supply considerably exceeded demand.

The Crispi era, 1887–1900. *Domestic policies.* On the death of Depretis in 1887 the Sicilian and former Mazzinian Francesco Crispi became prime minister and pursued a policy of administrative reforms at home and expansion abroad. His main domestic achievement was to extend the suffrage at local elections to all males over the age of 21 who paid five lire per annum in local taxes—that is, to 3.5 million people. This was a real blow to the local "notables" who had previously controlled local government. The larger councils (after 1896, all councils) were also permitted to choose their own mayors and were required to meet in public. The Crispi government also brought in a reasonably effective system of administrative law for the first time, through the provincial councils (*giuntas*) and the Council of State. The charities were reformed; the clergy was excluded from running them, and the funds were often diverted to more secular purposes. The minister of justice, Giuseppe Zanardelli, promulgated a new code of criminal law that abolished the death penalty and legalized strikes unless violence or intimidation occurred.

However, the most important act of Crispi's first government was the new tariff of 1887. It was a response to

The tariff of 1887

demands from northern steel and textile interests, from farmers (also mainly from the north) faced with imports of cheap American grain or Asian rice, and from social reformers eager to secure legislative measures that employers could afford. A duty of 50 lire per ton was placed on imported wheat by 1888, and later it went higher still; food prices rose sharply, provoking considerable unrest. Similar measures protected steel, shipbuilding, and textiles. Italy's largest trading partner was France, and the French retaliated against Italian goods. A "tariff war" began between the two countries and continued until 1898. Franco-Italian trade was more than halved, and entire sectors of Italian agriculture, including wine, silk, cattle, and olive oil, collapsed overnight as their markets were cut off. When excess food supplies drove all agricultural prices down, even grain growers failed to benefit from the new tariff. Moreover, the crisis helped to drag down many of Italy's banks, including one of the largest, the Banca Romana. Resulting inquiries revealed that the bank had made interest-free loans to leading politicians, including Crispi himself and former treasury minister Giovanni Giolitti, who was prime minister from May 1892 to November 1893. Politicians needed the money to finance their election expenses and to run or bribe newspapers. The Banca Romana scandal was the first of many famous Italian corruption scandals, and, like the others, it discredited the whole political system.

Colonialism. Crispi's colonial policy brought additional blows. The Italian settlement at Massawa soon led to conflict with the neighbouring Ethiopians, who in 1887 killed 500 Italian troops at Dogali. Peace was made at Wichale in 1889, and Crispi expanded the Italian possessions along the Red Sea to include most of present-day Eritrea and southern Somalia. In 1895 the Italians annexed a large portion of the province of Tigray, and war with Ethiopia began again. In March 1896 the Italian army was overwhelmed at the Battle of Adwa: about 5,000 Italian troops were killed. This disaster forced Crispi to resign and ended any further colonial adventures for some years. It was widely seen in Italy as a disgrace to the whole political system and to Italy's aspirations to great-power status; it would have to be avenged in the future.

Years of crisis. Economic hardship and political corruption at home, together with military failure abroad, provoked riots and uprisings throughout the country. In the early 1890s the "Sicilian *fasci*" (normally peasant leagues organized by urban socialists) led successful strikes and land occupations until Crispi, in January 1894, used the army to restore order. The *fasci*'s leaders were imprisoned, and the movement soon collapsed. At the same time, an anarchist insurrection in Lunigiana was also suppressed by martial law. Further riots occurred in 1898, mainly in towns, over the high price of bread and over civil liberties. At least 80 people were killed in Milan by the troops, whose commander was shortly afterward decorated by the king. Government repression also took the form of attempting to govern without Parliament (as Crispi did in 1895), of dissolving opposition associations and unions (as the government of Antonio Starabba, Marquis di Rudinì, did in 1897), and of attempting to push through restrictions on civil liberties by royal decree, without parliamentary approval (as both the di Rudinì and the Luigi Pelloux governments tried to do in 1898–99).

The Sicilian *fasci*

Repression soon led to a constitutional crisis. Conservative politicians, notably Baron Sidney Sonnino in 1897, argued that the Italian Parliament was corrupt and unfit to govern and that the king should provide strong executive rule, according to the letter of the 1848 Statuto (Statute). Most moderate Liberals rejected this argument. The campaign for constitutional government was led by Felice Cavallotti and the "Radical" group in Parliament, who in the 1890s strongly denounced bank scandals, tariff protectionism, colonial wars, and the Triple Alliance. The Radicals were a northern, anticlerical, moralistic group that denounced the corruption of the south (Crispi was the first southern prime minister), of the monarchy, and of the Roman establishment and strongly favoured wider civil liberties and army reform. In 1900, after months of bitter parliamentary dispute and obstructionism, Pelloux called a general election to resolve the constitutional issue; the Left triumphed; the Radicals won 34 seats, and their allies, the Republicans, won a further 28 out of a total of 557. (The two groups had had 51 seats between them in the previous Parliament.)

Furthermore, in 1892 a young Milanese lawyer, Filippo Turati, had helped to found the Italian Socialist Party among the various socialist and labour groups of northern and central Italy and Sicily. The new party was mainly social democratic, heavily influenced by the German model: it preached the class struggle and aspired to parliamentary representation and state socialism. Formally Marxist, it envisaged a long period of "evolution" before an eventual "revolutionary" transformation of society. Crispi dissolved the party in 1894, but it revived in the late 1890s and won 32 seats in 1900. While its deputies worked closely with the Radicals to secure constitutional liberties and social reforms, ordinary party members were often much more revolutionary in their aims. Other socialist organizations, such as trade unions and cooperatives, also grew in the 1890s and by 1900 were significant in the newly industrializing economy of northern Italy. They campaigned for concrete short-term gains on wages and working conditions and were usually more "reformist" than the party.

The Giolitti era, 1900–14. The elections of June 1900 marked the defeat of the Pelloux government and of attempts to impose illiberal laws. The following month King Umberto I was assassinated by an anarchist. The new king, Victor Emmanuel III, favoured a return to constitutional government, as did the governments led by Pelloux's successors, Giuseppe Saracco, Giuseppe Zanardelli, and Giovanni Giolitti, the last-named most often holding the office of prime minister between 1903 and 1914. His policy was to defuse popular discontent by social reforms and public works and to conciliate the major organized opposition groups in the country, the Socialists and the Roman Catholics.

Domestic policies. The social reforms passed in these years included laws prohibiting child labour, establishing a compulsory maternity fund and compulsory rest days, and limiting the working day of women to 11 hours. Central governments also subsidized municipal welfare schemes such as orphanages and old people's homes and encouraged municipal transport, housing, and water and sewage schemes—especially in northern Italy, where the municipalities could afford such innovations. Often these schemes were pioneered by Catholic- or Socialist-dominated local councils which entrusted the management to their "own" cooperatives; government approval of "municipal socialism" was much resented by local businessmen and taxpayers. Moreover, Giolitti's governments allowed trade unions to operate freely and generally avoided interfering in private-sector labour disputes—another source of middle-class resentment.

Giolitti enjoyed Radical support, and his governments often included Radical deputies. He also received the tacit support of moderate Socialist deputies and union leaders. Trade unionism grew rapidly in the new atmosphere after 1900, not only in industry but among the agricultural labourers of the Po Valley and Puglia; the various Socialist-led unions formed a General Confederation of Labour in 1906. Some unions depended heavily on public works schemes subsidized by government; others, such as the Federation of Agricultural Labourers (Federterra), relied on Giolitti's reform legislation favouring cooperatives and on contracts provided by Socialist councils. All the major Socialist institutions became reliant on government willingness not to repress them; in turn, they gave up any attempt to overthrow the government. However, the Socialist Party in the country was dominated by revolutionary views from 1904 to 1908 and was always more militant than its leaders. Moreover, there was also a powerful group of revolutionary syndicalists, who broke away from the Socialist Party in 1907 but still controlled many unions, especially in Liguria. This popular militancy ensured that Socialist deputies could not compromise too openly with Giolitti or accept posts in his governments.

General Confederation of Labour

The Catholics, too, had founded trade unions and workers' cooperatives, as well as mutual aid societies and rural banks, throughout northern Italy in the 1890s. These associations were particularly strong among the peasantry of Lombardy and Piedmont and among the largely female textile workers, and they, too, controlled many local councils. In 1897–98 most Catholic associations were dissolved by the di Rudinì government, but later governments permitted them to be refounded in return for tacit support against socialism. This support even became overt at parliamentary elections—in 1904 and 1909 the papal prohibition on Catholics voting (*non expedit*) was lifted in many constituencies, and Catholics were permitted to vote for Liberal candidates in order to keep Socialists out. The old "intransigents" of the Opera dei Congressi, deeply hostile to a united Italy, were replaced early in the century by a new generation of "clerico-moderate" leaders favoured by Pius X, who even dissolved the Opera dei Congressi in 1904 and brought the Catholic lay movement under the bishops. The clerico-moderates gave Giolitti their support, but they too could not enter government or even operate as a lay party independent of the bishops or the Vatican.

Economic developments. Giolitti's political dominance had rested on Italy's rapid economic growth after the mid-1890s. Industrial production probably doubled between 1896 and 1913. The tariff dispute with France was settled in 1898. Cotton remained the largest industry, but by 1914 Italy had also acquired—for military reasons—a large, protected steel industry, together with extensive shipbuilding yards in Liguria. The railways were nationalized in 1905, and this stimulated demand for rolling stock and engines. Hydroelectricity from the Alps provided cheap, renewable energy for the factories of the northern "industrial triangle" (Lombardy, Liguria, and Piedmont). Moreover, a major new industry—automobile production—developed, in which Italy did not have to compete against established interests elsewhere. Fiat, founded in Turin in 1899 by Giovanni Agnelli, soon became one of Europe's largest producers and exporters of automobiles and also made buses, trucks, airplanes, and military vehicles. The state finances were healthy, and the balance of payments was boosted by remittances from the enormous numbers of emigrants in Europe and the Americas.

Fiat

Agriculture was still the dominant sector of the economy, providing jobs for almost 60 percent of employed adults in 1911. It, too, enjoyed a boom, partly because of state-subsidized land reclamation and irrigation schemes (particularly in the Po Valley) and partly because of continued high tariff protection on grains, which gave ample incentive to produce more food on suitable land. Wheat production rose by about one-third in these years. In central Italy, sugar beet production, another heavily protected sector, stimulated a new refining industry.

Economic growth, however, was heavily concentrated in the north. The south languished, and income there was less than half that in the north. The southern economy was arguably linked more closely to northern Europe and South America (to which it exported wine, olive oil, fruit, and labour) than to northern Italy. Southern produce needed markets abroad, and the south was very badly hit by the tariff war with France. Moreover, the southerners were widely held, by the positivist school of anthropology fashionable in the 1890s and later, to be more criminal than northerners and even racially degenerate—an argument that lent racial overtones to the debate on "southern backwardness."

Southern politicians soon began demanding, and when in office securing, tax relief and development schemes, which provided, among other things, roads, schools, and irrigation. In 1897 the first "special law" provided Italy's poorest region, Sardinia, with cheaper credit and some funds for irrigation and reforestation; Sardinia's leading politician, Francesco Cocco Ortu, was minister of agriculture. Later laws extended similar or greater benefits to other regions and in 1906 to the entire south. In practice, the legislation had little impact, as World War I interrupted any progress. However, it was the first time that funds derived from taxes paid by the prosperous north were used by central government agencies to stimulate economic activity in the south—or, at least, to win votes for supporters of central politicians.

Emigration

Continuing southern poverty stimulated mass emigration from Sicily and the southern mainland, averaging more than 500,000 people per year from about 1901 onward and rising to 900,000 in 1913, mainly to North and South America. About half of the migrants to the New World returned later, bringing new values as well as new money. Some southerners crossed the Atlantic twice a year, moving to seasonal agricultural work in Argentina. In the north most emigration was seasonal, to other European countries; but many rural dwellers migrated within Italy to jobs in the expanding industrial cities. Migrants were usually young, male, unskilled, and illiterate. They tended to come from areas of small farms and peasant landownership, where there were no trade unions able to protect the rural poor and where people married late.

Health and education. The other major social changes in these years, apart from emigration, were those brought about by the decline in serious illnesses and in illiteracy. Improved water supplies and sewerage meant fewer cholera epidemics—though these still occurred at times, as at Barletta in 1910–12. Malaria, a major scourge of the rural south, declined sharply as quinine became widely available after 1900. Pellagra, a vitamin-deficiency disease endemic among the northern peasantry, rapidly declined as diets improved. By 1901, for the first time, a majority (51.3 percent) of Italians could read and write. Emigrants needed to be able to write home, and so they had an incentive to learn. In 1911 the primary schools were removed from municipal control—poor communes had not been able to build schools or to enforce attendance—and were henceforth run and financed by the central government. Millions of people now spoke Italian, having learned it in school or in the army or needing to use it as a lingua franca in the cities. A common language, common education, and common experience of military service had begun, by 1914, to "make Italians"—but religion, social class, and local loyalties still greatly divided them.

WORLD WAR I AND FASCISM

War and its aftermath. *Conduct of the war.* On Giolitti's resignation in March 1914 a new government was formed by the more conservative Antonio Salandra. In June came "Red Week," a period of widespread rioting throughout Romagna and the Marche, which was precipitated by the killing of three antimilitarist demonstrators at Ancona. When World War I broke out in August, the Salandra government stayed neutral and began to negotiate with both sides—a policy described by Foreign Minister Sidney Sonnino as "sacred egoism." The Austrians eventually agreed to grant Trentino to Italy, but the Entente (France, Britain, and Russia) made a more generous offer, promising Italy not only Trentino but also South Tyrol, Trieste, Gorizia, Istria, and northern Dalmatia. The Italians accepted this offer at the secret Treaty of London (April 1915) and joined the war against Austria-Hungary a month later, hoping for major territorial gains.

The negotiations, conducted by the foreign and prime ministers and a handful of diplomats, had been kept secret. The majority of deputies, meanwhile, favoured neutrality, as did former prime minister Giolitti, the major opposition groups (Catholics and Socialists), and most of the population. War, therefore, was supported only by the conservatives in government, by the Nationalist Association, a group formed in 1910 by Enrico Corradini and others to support Italian expansionism, by some Liberals who saw it as the culmination of the Risorgimento's fight for national unity, by Republicans and reformist Socialists who thought they were fighting for national liberation, and by some syndicalists and extremist Socialists—including Benito Mussolini, editor of the Socialist Party newspaper—who thought the war would bring about the overthrow of capitalism. Mussolini was soon expelled from the Socialist Party, but with help from the Entente he managed to found his own alternative, pro-war newspaper, *Il popolo d'Italia.* In April–May 1915 the government, helped by noisy demonstrations by pro-war activists, pushed through its war policy despite the opposition of the majority in Parliament and in the country. Neither Giolitti nor any other "neutralist" could form a government without renouncing

Supporters of the war

the Treaty of London, betraying Italy's new allies, and compromising the king.

In June 1916, after a series of military failures, the Salandra government resigned. The new prime minister was Paolo Boselli, who in turn resigned after a military disaster at Caporetto in October 1917, which enabled the Austrians to occupy much of the Veneto in 1917–18. The war was deeply unpopular both among the troops—mostly conscript peasants, who were undernourished and fighting for a cause few could understand—and among the civilian population back home, which included almost one million workers in arms factories who were also subject to military discipline. Serious bread riots took place among the industrial workers of Turin in August 1917, and the defeat at Caporetto was widely attributed to poor morale and "defeatism." After November 1917 a more "democratic" government under Vittorio Emanuele Orlando rallied the country to defend its frontiers and appointed a new army commander, Armando Diaz. Diaz made welfare concessions to the troops and fought a far more defensive campaign until October 1918, when, in the closing stages of the war, the Italians won a final, decisive victory at the Battle of Vittorio Veneto.

The cost of victory. Italy won the war, therefore, but at a huge cost: 600,000 dead and a legacy of bitterness and division. The victorious patriots and Nationalists now detested Parliament, where the Giolittian majority had never supported the war; the returning veterans also regarded Catholics and Socialists as defeatists and most industrial workers as "shirkers" (*imboscati*). These divisions greatly weakened the postwar political regime. Furthermore, the pro-war groups were themselves bitterly divided when the war ended. Should Italy, at the Paris Peace Conference (1919–20), try to secure the terms of the Treaty of London, as Foreign Minister Sonnino urged, or should it support U.S. President Woodrow Wilson and adhere to the "principle of nationality"—that is, be willing to accept less territory in the Adriatic region, as the Left Liberals and Republicans advocated? In the Treaty of Saint-Germain (1919), Italy was granted Trentino, Trieste, (the German-speaking) South Tyrol, and Istria. But Dalmatia was excluded, despite the Treaty of London, as was Fiume (now Rijeka), a Croatian port largely inhabited by Italian speakers, which Sonnino had also decided to claim; so, too, were any colonial territories in Africa or Asia and any claim on Albania. Nationalists therefore argued that Italy had been robbed of its rightful gains ("a mutilated victory"). Orlando resigned in June 1919. When the new government of the Radical leader Francesco Saverio Nitti was also unsuccessful in foreign affairs, the flamboyant poet Gabriele D'Annunzio led a group of volunteer "legionaries" to Fiume in September and captured the city himself. Fiume became a centre of nationalist agitation for more than a year, and D'Annunzio was dislodged only in December 1920, when Fiume became, briefly, an independent republic.

D'Annunzio's capture of Fiume

Economic and political crisis. Italy was faced with serious postwar economic problems. Wartime governments had printed money to pay for arms, and inflation became deep-rooted: by the end of 1920 the lira was worth only one-sixth of its 1913 value. Savings became worthless, and landowners found themselves receiving insignificant rents. Meanwhile the major arms and shipbuilding firms went bankrupt after the war for lack of government orders. Unemployment rose to two million, as returning soldiers searched for work. Peasants, organized by ex-servicemen's groups or Catholic leagues, seized land for themselves; agricultural labourers went on strike at harvest time. Trade unions, now operating again, pressed for higher wages, and strikes became routine, including those in the public services. Throughout September 1920 most of the northern Italian factories were occupied by workers staging sit-down strikes. The Socialist Party, dominated by its "maximalist" wing, proclaimed the need for immediate revolution, as had recently occurred in Russia. The postwar governments of Nitti (1919–20) and his successors Giolitti (1920–21), Ivanoe Bonomi (1921–22), and Luigi Facta (February–October 1922) were weak and could do little except urge industrialists and landowners to make concessions not only on pay but even on "control" of the workplace.

Diplomatic and economic failures soon undermined middle-class confidence in government, especially as Giolitti also imposed taxes on war profits. In 1919 proportional representation was introduced for parliamentary elections. The result, in the new Parliament elected in November 1919, was that the Socialists, with 30 percent of the vote, won 156 seats and the new (Catholic) Italian Popular Party, with more than 20 percent of the vote, 100 seats; these two parties dominated northern and central Italy. Giolitti had to bring the Popular Party into his government in 1920 and make many concessions to peasant interests, including giving guarantees to squatters and giving the Ministry of Agriculture to the Catholics. Furthermore, the two "subversive" parties won control of almost half the municipalities in the autumn of 1920, ensuring that Socialist or Catholic cooperatives would be given all local public works contracts. And in January 1921 the left wing of the Socialists split away to found the Italian Communist Party, which increased middle-class alarm.

The Fascist era. *The rise of Mussolini.* The political crisis of the postwar years provided an opportunity for militant, patriotic movements, including those of ex-servicemen and former assault troops, radical students, ex-syndicalists, and former pro-war agitators. D'Annunzio in Fiume led one such movement, but the ex-Socialist journalist Benito Mussolini soon became even more prominent, founding his Fasci di Combattimento ("Fighting Leagues") in Milan in March 1919. It was initially unsuccessful, but local Fascist groups were soon founded in Emilia and Tuscany and by autumn 1920 were busy not only breaking up strikes but also dismantling Socialist and Catholic labour unions and peasants' cooperatives and overthrowing newly elected local councils. Within a few months paramilitary Fascist squad leaders controlled most of the rural areas of central Italy. The Fascists had become a major political force, backed not only by landowners but also by many members of the urban middle class, including students, shopkeepers, and clerical workers. In May 1921, when Prime Minister Giolitti called new elections, 35 Fascists were elected to Parliament as part of a government "block" of 275 deputies. In October Mussolini abandoned republicanism, and in November he formed his movement into a proper political party, the National Fascist Party (PNF), which by this time was well-financed if ill-disciplined and extremely disparate: local bosses remained paramount in their areas. The Fascists also organized their own trade unions, the Fascist "syndicates," in order to replace Socialist or Catholic organizations, to provide mass membership, and to control labour.

Mussolini maneuvered brilliantly in the next few months, and the Liberal political establishment sought to conciliate him and the Fascist thugs. The police, the army, and much of the middle class sympathized with Fascist disruption of Socialist unions. Mussolini, as "duce" (leader) of fascism, gradually made himself indispensable in Rome. A major anti-Fascist protest strike, called by the Socialist-led Confederation of Labour in August 1922, soon collapsed, strengthening Mussolini's bargaining position even further. In October 1922 he organized a "March on Rome" by Fascist supporters. Prime Minister Facta asked the king to declare martial law, but Victor Emmanuel III eventually refused in order to avoid possible army disloyalty or even a possible civil war. Instead, he asked Mussolini to form a government, hoping to tame him by constitutional means.

Mussolini's March on Rome

Mussolini became prime minister, therefore, in a more or less constitutional manner. He was appointed by the king, and he headed a coalition government that included Nationalists, Liberals, and even (until April 1923) two Catholic ministers from the Popular Party. For 18 months he ruled through the usual government machinery and pursued a policy of "normalization." The squads were incorporated into an official Voluntary Militia for National Security. Ordinary middle-class job seekers flooded into the Fascist Party, making it more respectable and amenable; the Nationalists also merged their organization into it, bringing with them much respectable backing in the south. In 1923 the electoral law was changed once more, so that a group of parties with the largest vote—even if only 25 percent of the total—would receive an absolute majority of the seats. This enabled the Fascists to attract most of the old Liberal deputies into a "national alliance." In April 1924 elections were held under this system; the Fascist-dominated bloc won 64 percent of the votes and 374 seats, doing particularly well in the south. The opposition parties—by now including the Popular Party—remained divided but won a majority of the votes in northern Italy. The Socialists, indeed, had by this time split into three rival parties: the Communists, the Socialists, and the reformist Socialists. None of them had much influence. The Popular Party was being disowned by the Vatican, and its leader, Luigi Sturzo, resigned at the Vatican's request.

The end of constitutional rule. Mussolini's success as leader of a "normalizing" constitutional government with widespread support did not last long. When the new Parliament met, Giacomo Matteotti, leader of the reformist Socialists, denounced the recent elections as a sham and claimed there had been widespread intimidation of opposition voters. On June 10, 1924, Matteotti disappeared. He was later found to have been murdered by Fascist thugs led by the assistant to Mussolini's press office, Amerigo Dumini. A great crisis of confidence ensued. Mussolini was suspected of personal complicity and of having ordered the murder in order to eliminate a troublesome opponent. The press denounced the government, and the opposition parties walked out of Parliament. However, Mussolini still had a majority in Parliament, and the king backed him. But by autumn his Liberal supporters were drifting away, and in any case the "normalization" policy infuriated Fascist extremists in the country. They demanded a showdown, and Mussolini—who was too weak by this time to rule by constitutional means—had to agree. On Jan. 3, 1925, he made a famous speech in the Chamber of Deputies accepting responsibility for fascism and promising a tough crackdown on dissenters. The king made no move.

Mussolini's speech in the Chamber of Deputies

During the next two years Mussolini disbanded most of Italy's constitutional and conventional safeguards against government autocracy. Free speech and free association disappeared; opposition parties and unions were dissolved and their leaders forced into exile. At the local level, appointed podestas replaced elected mayors and councils. Freemasonry was outlawed—a real blow to most non-Catholic anti-Fascists. A Special Tribunal for the Defense of the State, run by militia and army officers, was set up to try anti-Fascist "subversives"; it imprisoned or sent to exile on remote islands thousands of political opponents, including the Communist leader Antonio Gramsci, and it imposed 31 death penalties. However, the repression was carried out essentially by the old state institutions, such as the police and army, not by Fascist bodies. The prefects—mostly still career civil servants—retained their traditional dominance over local government, and the new podesta was nearly always a landowner or retired army officer rather than a Fascist enthusiast. The Fascist party itself was soon swamped by more than a million job seekers and clerical workers, and thousands of the original Fascists were purged. The party, and the militia, soon had little to do except engage in propaganda and parades. The Fascist regime was mostly run by the traditional elites in the military and civilian bureaucracy, which were linked, as previously, to landowners and the court. That said, it was much more authoritarian and also much more nationalistic and interventionist than the Liberal governments had been.

Fascist indoctrination was never really successful, but the press was tightly censored, cinema newsreels were largely government propaganda, and the regime controlled the new radio broadcasting. It also ran semicompulsory Fascist youth movements, and new textbooks were imposed on the schools. The government also provided mass leisure activities, such as sports, concerts, and seaside holidays, which were genuinely popular.

Anti-Fascist movements. For a long time organized anti-Fascist movements remained weak, divided, and illegal and had no access to press or radio. The Communists were soon the most significant of these movements, as they had an underground organization and some Russian support and finance, but even they had 7,000 members at most and had great difficulty in spreading their propaganda in Italy. New anti-Fascist groups were founded occasionally, but the secret police soon cracked down on them. Apart from the Communists, only Justice and Liberty, an alliance of Republicans, Democrats, and reformist Socialists founded by Carlo Rosselli and others in 1929, managed to build up a clandestine organization in Italy. Most prominent anti-Fascists were in prison, in "confinement" on remote islands, or in exile and had little contact with Italian reality.

The only strong non-Fascist organization in the country was the Roman Catholic church. The Vatican implicitly supported Mussolini in the early years and was rewarded in February 1929 by the Lateran Treaty, which settled the "Roman question" at last. Vatican City became an independent state, Italy paid a large financial indemnity to the pope for taking over his pre-1870 lands, and a concordat granted the church many privileges in Italy, including recognition of church weddings as valid in civil law, religious education in secondary—as well as in primary—schools, and freedom for the lay Catholic organizations in Catholic Action. However, the government soon began curbing Catholic Action, seeing it as a front for anti-Fascist activity by former members of the Popular Party. The Catholic youth organizations were closed for a time in 1931. When they reopened, they had to avoid sports, but, even so, they grew considerably in the 1930s. They were a serious rival to the Fascist youth bodies and trained a new generation that often managed to avoid Fascist indoctrination. The 1929 concordat remained in force until the 1980s and was the legal basis for church dominance of Italian society after World War II. Anti-Fascist feeling became more widespread after the mid-1930s. Italy sent "volunteer" militiamen to fight on Francisco Franco's side in the Spanish Civil War (1936–39), but they were defeated in 1937 at the Battle of Guadalajara. Italy's increasingly close alliance with Adolf Hitler's Germany was also resented and feared, even by many Fascists. So, too, was the sudden decision to impose anti-Semitic laws in 1938: Jews were condemned as unpatriotic, excluded from government jobs, and forbidden to marry "Aryans." It had become clear that the Fascist government was likely to involve Italy in a disastrous European war, as indeed it did in 1940.

The Lateran Treaty

Economic policy. Fascist intervention in the economy was designed for prestige and to boost military strength. In the early years the Fascists compromised with the business establishment and rescued failing banks. However, in 1926 the lira was suddenly revalued for political reasons, and Italy suffered all the usual consequences of an overvalued currency. Exports fell sharply, unemployment rose, wages were frozen or even cut, and prices fell. The steel, electricity, and chemical industries expanded, for their markets were domestic, and they were helped by cheaper raw material imports; industries producing textiles, food, and vehicles, which were reliant on foreign markets, declined.

When the Great Depression came after 1929, these deflationary processes were accentuated. The leading banks, which had lent heavily to industry, had to be rescued in the early 1930s, as did many large industrial companies. Two new state-run holding companies, the Italian Industrial Finance Institute (Istituto Mobiliare Italiano; IMI) and the Institute for Industrial Reconstruction (Istituto per la Ricostruzione Industriale; IRI), were set up to bail out failing firms and to provide capital for new industrial investment; they also provided trained managers and effective financial supervision. Italy thus acquired a huge, state-owned industrial sector, which was especially important in banking, steel, shipping, armaments, and the supply of hydroelectricity. However, these firms were not nationalized; they operated in the market as private companies and still had many private shareholders. In the long term they gave Italy a modern infrastructure—including roads and cheap energy—a sounder financial sector, and some efficient modern industries in expanding sectors such as chemicals and synthetic fibres. Most industrial development, and most workers, remained in northern Italy, although by this time large steelmaking and shipbuilding plants had been started at Naples and Taranto.

The Fascist syndicates

After October 1925 the Fascist syndicates, or trade unions, were the sole recognized negotiators for workers' interests. Strikes and lockouts became illegal, and wages fell between 1927 and 1934, but the syndicates had considerable political influence. They secured a shorter working week (40 hours in November 1934), higher welfare benefits (such as family allowances, also introduced in 1934), and public works schemes, and they also helped run leisure and social activities. Most industrial jobs were protected throughout the depression, and the syndicates remained fairly popular until World War II. In 1934 the Fascists also set up "corporations"—mixed bodies of syndicalists and employers—to decide labour disputes and supervise wage settlements. Despite much rhetoric and propaganda about them, they had little impact in practice and virtually none on industrial management or economic policy making.

In agricultural policy the government aimed at self-sufficiency by encouraging grain production ("the battle for wheat"). A high tariff was reimposed on imported wheat, and grain prices were kept artificially high. Production rose sharply as northern farmers used more chemical fertilizers. In much of the south the climate was unsuitable for growing wheat, but vineyards and olive groves were nonetheless plowed up. The real beneficiaries of this policy were the large farmers of the Po Valley and of the southern latifundia. These men also benefited most from the government's land-reclamation schemes, forming their own consortia and receiving government money to drain or irrigate their own land. Moreover, during the depression they could buy land cheaply from the smaller landowners because many of the peasants who had acquired land during and after World War I were forced to sell after 1926.

After Italy's invasion of Ethiopia in 1935–36, the Italian economy was subjected to sanctions by the League of Nations. This led to a more extensive drive for self-sufficiency ("autarchy"); imports were replaced where possible by native products, and most exports were diverted to Germany and Switzerland or to Africa. Ethiopia, once conquered, became a vast drain on resources. Government intervention and licensing became more marked, with official cartels and quasi-monopolies encouraged; resources were shifted, from above, to heavy industry and armaments. All this led to budget deficits, big tax increases, and capital levies, which were hugely resented because they were mainly needed to pay for wars in Africa and Spain. Resented, too, was the obvious corruption of the Fascist governing clique, without whose permits—available at a price—nothing could be done. The various conservative groups, including those in the army, the civil service, the law, and the church, which in the mid-1920s had looked to Fascism to protect their interests, realized by the late 1930s that Fascism was unreliable and withdrew their support.

Declining emigration

American restrictions, European recession, and Fascist economic nationalism combined to curtail emigration drastically in the 1930s, from more than 600,000 people per annum before 1914 to fewer than 50,000 per annum. The closing of emigration outlets hit the south particularly badly. As they could not go abroad, rural Italians moved to the cities: Rome doubled in size between 1921 and 1940, and northern cities became full of southern immigrants. Moreover, government policy encouraged population growth by providing tax incentives to have children and excluding the childless from public jobs. Admittedly, all this had little effect before 1937: Italians married later than ever and had fewer children than previously, so much so that in several northern and central regions the birth rate dropped below replacement level in the 1930s.

Foreign policy. Fascist foreign policy became more expansionist as time went on. In particular, Mussolini aimed at acquiring territory in Africa and in the Mediterranean, which he termed "mare nostrum" ("our sea"). Even in 1923, in his first year in office, he briefly invaded the Greek island of Corfu to avenge the murder of four Italian nationals forming part of an international boundary delegation. During the next decade he played the European statesman, and in 1924 he reached an agreement with Yugoslavia that gave Fiume to Italy. He also continued to strengthen the Italian hold on Libya, to build up the armed forces, and to plan further expansion in Africa—particularly in Ethiopia, where the defeat at Adwa in 1896 still needed to be avenged. In October 1935 Italy finally invaded Ethiopia and by May 1936 had conquered the country. However, the war antagonized the British and French governments, led to sanctions by the League of Nations, and isolated Italy diplomatically. Mussolini moved into Hitler's orbit, hoping that German backing would frighten the British and French into granting further concessions to Italy. The policy was unsuccessful: no more territory was acquired in Africa. Furthermore, Italy became a junior partner in the "Rome-Berlin Axis," and in 1938 Mussolini had to accept Hitler's annexation of Austria, bringing the German Reich right up to the Italian border. In May 1939 Mussolini entered a formal military alliance with Hitler, the "Pact of Steel," which further reduced his scope for maneuvering. Even so, when the Germans unexpectedly invaded Poland in September 1939, Mussolini insisted on remaining neutral.

World War II. *Military disaster.* Only in June 1940, when France was about to fall and the war seemed virtually over, did Italy join the war on Germany's side, still hoping for territorial spoils. Italy's initial attack on the French Alps in June 1940 was soon cut short by the Franco-German armistice. The real war for Italy began only in October, when Mussolini attacked Greece in a disastrous campaign that obliged the Germans, in 1941, to rescue the Italian forces and take over Greece themselves. The Germans also had to lend support in the hard-fought campaigns of North Africa, where eventually the decisive second battle of el-Alamein (October 1942) destroyed the Italian position and led to the surrender of the whole of Italy's North African forces in May 1943. Meanwhile, the Italians had lost their extensive empire in eastern Africa, including Ethiopia, early in 1941; and their 250,000 troops in Russia, sent to help the German invaders, suffered untold hardships. In short, the war was an almost unrelieved succession of military disasters. The main reason was poor equipment. Italy had few tanks or antitank guns; clothing, food, vehicles, and fuel were all scarce; and supplies could not safely be taken to North Africa or Russia. Italian factories could not produce weapons without steel, coal, or oil, and, even when raw materials were available, production was limited because the northern Italian factories were subject to heavy Allied bombing, especially in 1942–43.

Mounting opposition to the war

Bombing, indeed, was one of the causes of the first major strikes since 1925. In March 1943 the leading factories in Milan and Turin stopped work in order to secure evacuation allowances for workers' families. By this time civilian morale was clearly very low, food shortages were endemic, and hundreds of thousands of people had fled to the countryside. Government propaganda was ineffective, and Italians could easily hear more accurate news on Radio Vatican or even Radio London. In Friuli–Venezia Giulia, as in Italian-occupied Slovenia and Croatia, the local Slav population supported armed resistance movements, and anti-Italian terrorism was widespread. The anti-Fascist movements cautiously revived in 1942–43. The Communists helped to organize strikes, the leading Roman Catholics formed a Christian Democratic Party in 1943, and a new Party of Action was founded in January 1943, mainly by republicans and Radicals. By this time most of the leading clandestine parties were more willing to work together to overthrow fascism; in April 1943 they signed an agreement to do so.

By the summer of 1943 the Italian position was hopeless. Northern and eastern Africa had been lost, the northern Italian cities were being regularly bombed, war production was minimal, and morale had collapsed. Court circles began sounding out Allied terms, which of course included the ouster of Mussolini. In July 1943 the Allies invaded Sicily, and within a few weeks they controlled the island. On July 24–25 the Fascist Grand Council met in Rome and passed a motion asking the king to resume his full constitutional powers—that is, to dismiss Mussolini. The king did so the same day and installed Marshal Pietro Badoglio as prime minister. The army took over the key positions in Rome, the duce was arrested, and the main Fascist institutions were dissolved.

End of the regime. Badoglio assured the Germans that the war would continue, but he also negotiated an armistice with the Allies, which was declared in September 1943. The Germans immediately took over Rome; they had already seized most of central and northern Italy. The Italian army disintegrated, and the king and his government fled south. Italy became a war zone: for 18 months the Allies and Germans fought each other up the peninsula, devastating the land. The Allies took Naples in October 1943 but reached Rome only in June 1944, Florence in August, and the northern cities in April 1945.

The Allies ruled the south, and Badoglio's government had very little influence on events. When Rome was liberated, the leading anti-Fascist parties formed a new "government." Meanwhile, the Germans had rescued Mussolini from his mountain prison and made him the puppet ruler of the northern "Italian Social Republic," a last-ditch Fascist regime based in Salò on Lake Garda. He, too, had little influence on events. At the end of April 1945 he fled toward Switzerland but was captured by Communist partisans at Dongo and shot; his body was then brought back for public display in Milan.

Mussolini and the Italian Social Republic

The partisans. After September 1943 partisan groups were active throughout northern and much of central Italy. Often they were former soldiers; many were young men fleeing Mussolini's attempts to conscript them; others were urban evacuees. They were most active in summer in the hills, where peasants usually supported them, and they tied down thousands of German troops. There were also major strikes in industrial areas sabotaging war production. The Communist Party led the largest group of partisans (at least 50,000 by summer 1944). Success in the Resistance transformed the Communists into a major force in postwar Italian politics. Other parties fielded tens of thousands of additional fighters. Partisans of different political persuasions normally worked together in local Committees of National Liberation, which coordinated strategy, administered liberated areas, and organized the uprisings in the northern cities, including Milan in April 1945, which fell to the partisans before Allied troops arrived. They thus laid the basis for postwar political collaboration. (M.Cl.)

Italy since 1945

THE POSTWAR DECADES

Birth of the Italian republic. With the German surrender in May 1945, the anti-Fascist parties formed a government dominated by northerners and led by the Resistance hero Ferruccio Parri. Up to 15,000 Fascists were purged or killed. The purges caused much alarm among public-sector workers, virtually all of whose jobs had been contingent on Fascist Party membership. In reality, the purges were short-lived and superficial, and there was considerable continuity in many areas, including the judiciary, the police force, and the legal code. In November 1945 Parri was replaced by the Christian Democratic leader, Alcide De Gasperi, who formed a more moderate—and southern—interparty government. It ended purges, returned large industrial firms to their previous owners, and replaced partisan administrators with state officials. In May 1946 King Victor Emmanuel III abdicated. His son briefly became King Umberto II, but the royal family had to leave the country a month later when a majority voted in favour of a republic.

At the same time, a constitutional convention was elected by universal suffrage, including women for the first time. Most of the votes and seats went to the three largest parties, the Socialists, the Communists, and especially the Christian Democrats, who began their postwar dominance as the most powerful party. Promulgated in 1948, the constitution established a parliamentary system of government, with two houses (Chamber of Deputies and Senate) elected by proportional representation. It provided for civil and political rights, an independent judiciary, a constitutional court, and the right of referendum. Elected by Parliament, the president was to have few real powers. The 1929 Lateran Treaty with the church was confirmed. Regional governments were established in linguistically or ethnically distinct outlying zones—Sicily, Sardinia, Valle d'Aosta, and Trentino–Alto Adige (including South Tirol). In short, the constitution was an "anti-Fascist" document, providing for weak governments and individual liberty—exactly the opposite of what Mussolini had attempted.

The Cold War political order. In 1947 De Gasperi received $150 million in aid from the United States. He had excluded the Communists and the Socialists from his government both to placate the conservative South and to ensure the much-needed American aid. As parliamentary elections approached, U.S. Secretary of State George Marshall threatened that aid would be cancelled if the Communists and Socialists came to power. During the first postwar parliamentary election campaign in 1948, the United States provided huge backing for the Christian Democrats and their Liberal, Social Democratic, and Republican partners. The Christian Democrats, also backed by the church, won more than half the seats in Parliament. The Communist-Socialist alliance won nearly one-third. In July an assassination attempt on the popular Communist leader, Palmiro Togliatti, sparked a wave of Communist unrest. Although Togliatti called for calm and the unrest soon subsided, the Christian Democrats accused the Communists of seeking to overthrow the government.

Italian politics set for the next 40 years into its "Cold War" mold. The Christian Democrats, with American support, controlled the government and shared power and patronage with their smaller pro-Western coalition partners. They appealed to peasant landowners and women, and generally won about two-fifths of the vote. The Socialist Party ended its alliance with the Communists in the late 1950s and began to cooperate with the Christian Democrats. Beginning in 1963, the Socialists joined "centre-left" coalition governments, acquiring control of key ministries and public-sector enterprises. The Communists, excluded from the central government, won about one-third of the vote, were strongest among industrial and farm workers, ran local governments in central Italy, and controlled the major trade unions.

Parties and party factions. All of the major Italian parties had large memberships recruited from organizations such as Catholic Action, cooperatives, and trade unions, which often provided tangible benefits—jobs, disability pensions, and cheap holidays—for their members. Distinct subcultures, including newspapers, bars, theatres, and schools, grew up around each major party. Most parties comprised organized "factions," each with its own leaders, deputies, regional or ideological base, sources of finance, and journals. Within each party, these factions contended for power and for control of public-sector firms and agencies to secure resources and jobs for supporters. Governments between 1945 and 1994 were short-lived, with an average life of 11 months, partly because secret ballots in Parliament enabled dissatisfied ruling-party factions to bring down governments. Disgruntled factions could then demand posts (and access to patronage) in a new government. However, instability was more apparent than real—top politicians often held key government posts semipermanently.

To assure their reelection, deputies needed to win favours for constituents. The state sector of the economy, already large in 1945, expanded after the war, and new jobs often went to party members. In turn, state firms financed parties or factions. In many areas, especially in the south, party-controlled agencies dominated economic and social life. Local government could rarely operate without favours and finance from central, party-controlled agencies. Because of the strength of the Christian Democratic Party in the south, southerners increasingly predominated in government posts, even in the north. This extensive, and expensive, system of patronage was known as *partitocrazia* ("partyocracy"). Although this system was obviously corrupt, it commanded a broad public consensus.

Foreign policy. The Cold War political system had one major advantage: Italian foreign policy ceased to be adventurous. De Gasperi had to accept the harsh Treaty of Paris in 1947, in which Italy gave up all African colonies and relinquished some Alpine territories to France and the

Dodecanese islands (held since 1912) to Greece. The Trieste issue was settled in 1954, when the city became part of Italy, while the area to the south joined Yugoslavia. Italy joined the North Atlantic Treaty Organization (NATO) and the Western alliance. NATO—in effect, the United States—guaranteed the country's political stability and security. Italy also was a founding member in 1957 of the European Economic Community (EEC; now part of the European Union [EU]).

The economic miracle. *Industrial growth.* The republic enjoyed economic success for many years. Initial American support helped to rebuild basic industries, including steel. The government embarked on an ambitious reconstruction program, and prewar industrial production levels were regained by 1948. Italy became fully integrated into European trade and was increasingly active in Middle Eastern oil extraction. Until 1964 the country enjoyed an "economic miracle," with growth rates of more than 8 percent annually. Its most prominent industries, still in the northwestern industrial triangle, produced fashionable clothing (especially shoes), typewriters, appliances, furniture, plastics, synthetic fibres, motor scooters, and cars. Italian firms combined elegant design and inexpensive production techniques. A network of superhighways was constructed across Italy. In less than two decades, a largely agricultural backwater became one of the world's most dynamic industrial nations. Economic success gave politicians additional resources for political patronage.

The postwar recovery and subsequent expansion benefited from a stable currency from 1948 onward and from cheap access to Middle Eastern oil. Rural migrants flooding the cities provided inexpensive labour, trade unions were weak and politically divided until the late 1960s, regulatory agencies were even weaker, and taxes were low and easily avoidable. All this encouraged investment, especially with cheap credit provided by state-owned banks. The IRI, founded under Mussolini in 1933, continued to dominate sectors such as heavy industry, telephone service, air transport, and highway construction. The economic miracle, therefore, did not rest on market principles alone; the government played a vital role.

Land reform. In agriculture the major postwar change was the land reform laws of 1950, which enabled state agencies to expropriate large, badly cultivated estates—mostly in southern or central Italy—for improvement and sale to new peasant owners. However, most large landowners escaped expropriation, and only 117,000 families actually acquired farms. Landless peasants moved abroad or to the cities instead. However, land reform enabled reform agencies—run by politicians in Rome—to dominate many rural areas, controlling land allocation, loans, and improvement grants. Christian Democratic political patronage thus replaced some of the traditional local power of landowners. Meanwhile, mechanization gradually eliminated many traditional jobs in the Italian countryside. Small, well-managed farms prospered, in part through EEC subsidies, and rural towns grew.

The south. The major economic problem was still the relatively underdeveloped and impoverished south, where there was little industry. Because land reform actually transferred little land to the peasantry, many of the south's inequities remained. A special Southern Development Fund (Cassa per il Mezzogiorno), established in 1950, financed roads, schools, electrification, and water provision. After 1957 it also began to invest in industrial and infrastructural development. The policy produced several large, capital-intensive heavy industry plants that produced uncompetitive goods, employed little local labour, and caused serious environmental harm. Although a few areas did take off (notably the Puglian coast north of Bari), the industrialization policy soon faced widespread criticism. Northerners resented the cost, and southerners could see little benefit. Although funding gradually shifted to smaller projects, the fund failed to industrialize southern Italy, where unemployment remained well above the national average and wages well below it.

The south did benefit, however, from some of the fund's investments, such as decent roads, clean water, improved health services, secondary schools, and malaria eradication. It also received state social spending and agricultural subsidies—often provided by politicians seeking votes. The south also benefited from emigrants' remittances. More than three million people left the south between 1955 and 1970. Some rural areas became seriously depopulated, while Rome and many northern cities nearly doubled in size, as migrants crowded into bleak housing estates or improvised shantytowns on the outskirts. (M.Cl./J.Fo.)

The effects of southern emigration

ITALY SINCE THE 1960S

Beginning in the 1960s, Italy matured into one of the most economically and socially advanced countries of the world. Demographic, economic, and cultural changes transformed Italian daily life and fueled social unrest in the 1960s and '70s. As economic growth slowed during the 1970s and '80s, migration from the south decreased and even reversed. After the Cold War ended in 1989, pressures for political and economic reform and European economic unification exposed Italy to a new range of challenges.

Demographic and social change. Population growth in Italy had slowed dramatically by the 1960s. The birth rate in the north and centre fell below replacement level in the 1970s; even in the south, birth rates fell sharply after 1964. By the 1990s Italy had one of the lowest birth rates of any industrial country. Births declined for many reasons. Contraception became readily available after 1971, and most Italians were now urbanites who no longer needed lots of children to help till the soil. Moreover, women were now better educated. After female attendance at secondary schools rose in the 1960s, women could readily find jobs that gave them financial independence from men and alternatives to lives as homemakers and mothers. During the 1970s, antiquated provisions of family law were repealed, and new laws permitting divorce and abortion were confirmed in national referendums, despite church opposition. Meanwhile, more people chose not to marry.

Regular church attendance fell sharply, from more than two-thirds of the population in the mid-1950s to less than one-third in the 1980s. Furthermore, many church-controlled charities, especially at the local level, were taken over by regional governments in the 1970s and run as part of the state welfare system. By the 1980s Italy was indeed "de-Christianized," as Pope John Paul II said. In 1984, under a new concordat between the Vatican and (significantly) the government of Socialist Bettino Craxi, Roman Catholicism ceased to be the state religion, religious instruction in schools became voluntary, and the state stopped funding priests' salaries.

Economic stagnation and labour militancy. After 1963, when the Socialist Party entered government, trade unions became more powerful. Inflation began to rise as governments printed money to pay for higher wages and benefits. Many firms had to be rescued by the IRI at public expense, the balance of payments deteriorated, and the official economy began to slow down, although the black-market economy of self-employed artisans, among others, continued to flourish.

This economic slump led to the "hot autumn" of 1969, a season of strikes, factory occupations, and mass demonstrations throughout northern Italy. Most stoppages were unofficial, led by workers' factory committees or militant leftist groups rather than by the (party-linked) trade unions. The protests concerned not only pay and work-related matters but also conditions outside the factory, such as housing, transport, and pensions, and they formed part of a larger wave of political and student protest, including opposition to the Vietnam War.

The stoppages forced employers to grant large pay raises. In 1970, legislation—the Statute of the Workers—ratified these developments. In 1975 most pay scales were indexed to inflation on a quarterly basis, thus guaranteeing the big pay raises of the previous few years. Jobs, too, were virtually guaranteed in the official economy, and the firing of workers became extremely difficult in many sectors.

Deterioration of the economy

Labour militancy continued into the 1970s, often led by unofficial "autonomous" unions. Many firms therefore chose to restructure into smaller units employing part-time or unofficial workers on piece rates who could be dismissed easily and did not enjoy guaranteed wages. Unemployment rose sharply, especially among the young. Inflation continued, aggravated by the oil price rises of the 1970s. The budget deficit became permanent, averaging about one-tenth of gross domestic product, a record for any industrial country. The lira lost more than half of its value between 1973 and 1982.

Student protest and social movements. Student protests in Italy also spread in the late 1960s and continued through the 1970s. Universities were occupied, and alternative life-styles began to dominate youth culture. Students challenged the church, the Communist Party, consumer society, and the traditional power of the family. However, after an initial creative and democratic phase, the movement fell under the shadow of small, violent sectarian factions.

The feminist movement also exploded in Italy in the mid-1970s, later than in most other Western countries. Feminists challenged rigid Catholic morals, a legal system that gave women little defense against male oppression, and male dominance of politics, even on the left. Feminist agitation helped mobilize support for the referendums of the 1970s on divorce and abortion.

Even the church began to open up to social and cultural change. The Second Vatican Council (1962–65), called by the reformist Pope John XXIII and implemented by his successor Pope Paul VI, provided a framework for the partial liberalization and democratization of the church. The process of liberal reform and hopes for a transformation of the church, however, declined with the succession of the more conservative Pope John Paul II in 1978.

Terrorism. When economic, social, and political stability suddenly collapsed after 1969, one of the most alarming results was terrorism. Initially, neofascist groups began planting bombs and derailing trains as part of a "strategy of tension" to undermine the labour advances of 1969–72 and encourage a right-wing coup. Initial police suspicion fell upon the far left, especially the anarchists. Meanwhile, evidence emerged—which the police had ignored—that neofascists had planted the bombs with the support of sectors of the Italian secret services. The "strategy of tension" continued until 1984.

By the mid-1970s, meanwhile, left-wing terrorism had begun to attract politically disaffected young people. Hundreds of former students and unemployed workers in many small groups kidnapped and sometimes killed officials and journalists. The best-known group, the Red Brigades, kidnapped and murdered former prime minister Aldo Moro in 1978. In response, the police gained special powers, the courts gave captured terrorists incentives to provide evidence, and by 1982 terrorism was greatly reduced.

The Red Brigades

Politics in the 1970s and '80s. The political system survived with the assistance of the Communists. In the face of terrorism and economic crisis, the Communists, led by Enrico Berlinguer, adopted a policy in 1973 that he called the "historic compromise." This entailed more or less formal alliances with the Christian Democrats for the good of the nation. During this period, patronage networks spread across the political system. However, Berlinguer's "historic compromise" alienated many Communist supporters, and Communist cooperation ended in 1979.

Governments in the 1980s were usually four- or five-party coalitions in which the smaller parties played an increased role. The Christian Democrats, weakened by secularization, factional disputes, and scandals, lost support steadily during the 1980s and ceded the premiership for the first time since 1945. Craxi, a Socialist, was prime minister from 1983 to 1987.

Socialists, in fact, secured many key posts in the 1980s, not only in government but also in public-sector firms. Disputes among and within the leading parties over the allocation of jobs and resources became more prolonged and often paralyzed effective government. Corruption scandals increased, the public debt rose to unsustainable levels, and popular resentment of *partitocrazia* mounted. The system could no longer deliver the patronage that had sustained it, and the state-dominated economy was falling behind those of other European countries.

Regional government. During the 1970s elected regional assemblies and governments, which had previously existed only in a few outlying regions, were finally set up throughout Italy, as the constitution had required. They acquired extensive powers, especially over agriculture, health, social welfare, and the environment. In 1978 many national agencies were dissolved and their powers allocated to the regions. In 1984 even the Southern Development Fund was abolished and its planning and investment powers redistributed. The allocation of power to the regions decreased the influence of central politicians. In the north politicians became more conscious of regional interests and more intent on autonomy from Rome. In the south, however, continuing poverty sustained a need for subsidies from the central government.

The effects of regionalism

The economy in the 1980s. Economic growth revived in the mid-1980s, once terrorism had ended and the 1979 oil crisis had subsided. The long season of protest that had begun in 1969 finally ended. As employers shifted production to smaller, nonunion units, the power of trade unions decreased. Big industry began to decline, especially in the industrial northwest, and a 1985 referendum reduced the indexing of wages. Nonetheless, northern Italy prospered in the middle and late 1980s, helped by the low price of oil, and people spoke of a "new economic miracle."

The Italian economy began to develop along new lines. Small businesses flourished in central and northeastern Italy—collectively known as the "third Italy," alongside the northwest, with its older industries and financial centres, and the less-developed south. These firms mainly produced quality goods for export. New industrial districts in these regions specialized in particular products, from faucets to ties. New sectors, such as fashion and design, began to replace traditional industries in the northern cities. Milan became one of the world's fashion capitals, bringing in billions of lire in business and advertising. The introduction of private television in the 1970s created a new industry and boosted the fortunes of dynamic entrepreneur Silvio Berlusconi.

Still, serious problems persisted. Budget deficits remained large and intractable. The economy continued to depend heavily on "unofficial" work done by casual workers in small firms and service industries (the so-called black-market economy). Most of the south, moreover, did not participate fully in the economic recovery. It did, however, maintain a thriving black-market economy supported partly by organized crime activity. As emigration diminished and mass education expanded, living standards began to rise, although they remained well behind those in the more affluent north.

The southern economy

Public services remained an economic and political quagmire and a target of growing popular outrage. Despite centres of excellence, public agencies were top-heavy with bureaucracy, inefficient, and corrupt. Not only did Italians pay more taxes than most other western Europeans, but the services they received were often comparable to those of Third World countries. Daily dealings with the state brought frustration and anger to many Italians.

The fight against organized crime. Organized crime dominated whole regions of the south politically, socially, and economically by the 1980s and thrived in cooperation with corrupt officials who allowed the criminals to siphon off public resources and skirt the law in return for political support. However, during the mid-1980s the state and civil society began to attack criminal organizations. Key figures began to turn state's evidence in defiance of their organizations' codes of silence, and the state succeeded in stemming organized crime. Corruption investigations in the early 1990s led to the prosecution of previously immune political figures. Seven-time prime minister Giulio Andreotti was charged with collusion with the Mafia in 1993, a move that shook the political system to its foundations, although Andreotti was later absolved. The south began to stabilize, although nobody believed that organized crime had been entirely defeated.

Italy in the 1990s. *Collapse of the Cold War order.* The end of the Cold War in 1989 and growing economic globalization ended the pattern that had marked Italian politics since the 1940s. Most Italians had come to resent the corrupt and costly system of patronage and the large state economic sector that hampered Italy's competitiveness and tarnished its political culture. With the bankruptcy of Cold War ideology and *partitocrazia*, the party system itself began to appear outdated, and shifting alliances replaced solid party blocs.

The end of the Cold War had an immediate impact on the two biggest parties. The Christian Democrats' ongoing scandals and their loss of anticommunist appeal further reduced their popularity. Meanwhile, the Communist Party adopted a more moderate program and, in 1991, a new name: the Democratic Party of the Left (Partito Democratico della Sinistra; PDS). The PDS vote declined, but it remained the leading centre-left party.

Operation Clean Hands

Public protests against political corruption had little effect until 1992, when investigators uncovered a series of bribery scandals. Many leading politicians, civil servants, and prominent businessmen were arrested and imprisoned. Nearly all of the parties were involved, but the Christian Democrats and the Socialists were at the heart of the system. Craxi, the former prime minister, was convicted on multiple charges and escaped imprisonment only by fleeing to Tunisia, where he died in 2000. By 1993 more than 200 deputies and several former ministers were prosecuted in a series of closely followed trials.

Apart from the PDS, whose role in the corruption was limited, the main political parties dissolved in disgrace, some to reappear under new names and with new leaders. Christian Democrats formed the Italian Popular Party (Partito Popolare Italiano; PPI). Neofascists, who had remained largely outside of the system of corruption, founded the new National Alliance (Alleanza Nazionale; AN). The Socialists, so important to the political system since the 1960s, became irrelevant. It was an unprecedented reshaping of an entire political system.

In 1993 voters approved referendums, later ratified by Parliament, to alter the electoral law so that three-quarters of the deputies and senators would be elected from single-member constituencies rather than proportionally. Socialist prime minister Giuliano Amato (1992–93), whose government had been rocked by the corruption scandal, resigned shortly after the passage of the referendum, and President Oscar Luigi Scalfaro named a banker, Carlo Azeglio Ciampi, as prime minister to implement the electoral reforms and stabilize the economy. The collapse of the existing party system and elections under the new law the following year marked the end of *partitocrazia* and the beginning of a new political order.

Economic strength. Economic problems increased sharply after 1991, as a global recession touched Italy. In an effort to reduce the budget deficit, Amato's government abolished the indexing of wages, decreased social spending, and drew up a program of gradual privatization.

During the mid-1990s, the economy resumed the strong growth of the previous decade, led by the flourishing design and manufacturing small-business sectors of the centre and north. Living standards throughout Italy rose to the levels of the most advanced economies, although large pockets of poverty remained, particularly in parts of the south and on the bleak peripheries of the northern cities.

Most Italians enthusiastically supported the EU, and Italy participated in the move toward European integration. To meet the strict requirements of the common European currency, the government of Romano Prodi (1996–98) passed austerity budgets that dramatically reduced Italy's chronic budget deficits. Under the Prodi government, privatization began in earnest, and inflation fell to record lows. This fiscal discipline allowed Italy to adopt the common European currency, or euro, in 1999.

Immigration. The economic growth that had begun in the 1980s drew growing numbers of immigrants to Italy. Most worked in the service sector or in small-time street trading. Immigrants had great difficulty finding housing, and the Italian state did little to help. Nonetheless, immigration increased in the 1990s. By 2000, there were over a million immigrants in Italy, many of whom lacked residence permits. Racism emerged in Italian society and politics, and immigrants were stereotyped as criminals, just as southern Italians had been in the 1960s. However, their cheap labour was essential to the Italian economy.

In 1993 Italy ratified the Schengen Treaty, which eliminated passport controls between its European member states and mandated rigorous controls for persons arriving from nonmember states. Italy implemented these controls and joined the Schengen zone in 1997. Because of its position at the edge of prosperous western Europe, Italy played a frontier role; every day immigrants attempted the perilous sea crossings from Albania or North Africa, with Italian authorities trying to stop them.

A new political landscape. Meanwhile, a new political pattern began to emerge. In the north, regional "leagues" had developed in the early 1980s to protest corruption in Rome, high taxes, poor public services, organized crime, and immigration. These leagues united in 1991 to form the Northern League (Lega Nord; LN) and won growing northern support. In the mid-1990s, the league called for a northern republic that would have autonomy from Italy except for defense, foreign affairs, and monetary policy.

The Northern League

The Northern League filled part of the vacuum created by the collapse of the major parties, as did the "post-Fascist" AN, which in Rome and the mainland south became the voice of Italian patriotism and of continuing state subsidies and economic intervention. Above all, the political vacuum was filled in 1994 by Silvio Berlusconi, who controlled much of the media and the successful A.C. Milan football (soccer) club. Berlusconi hastily founded a party, Forza Italia ("Go Italy"), with a message of populist anticommunism and formed an electoral alliance with the LN and the AN. This loose right-wing coalition won a slim majority of seats in the 1994 parliamentary election, the first held under the new electoral law.

Berlusconi, who became prime minister, had pledged to cut taxes, lower public spending, decentralize government, and generate a million new jobs. However, the new government was just as faction-ridden as previous ones. Berlusconi himself faced allegations of corruption, and his attempt to halt the anticorruption investigations and to control the state-owned media (which competed with his own holdings) failed miserably. The LN broke with Berlusconi and brought down the government. In 1995, President Scalfaro once again named a former banker, Lamberto Dini, to form a government of nonpolitical "technocrats," supported by the NL and the left-wing parties in Parliament. After the collapse of Italy's party system, powers that the parties previously had exercised—such as the decision to dissolve Parliament—rested increasingly with the president.

In 1995 Prodi, a former Christian Democrat, proposed a new centre-left alliance, known as the Olive Tree (L'Ulivo). He promised to prepare Italy for the common European currency and to reform the bureaucracy, and he won the support of the PDS and the other major left and centre parties. The division on the right in 1996 allowed the Olive Tree to win a surprise victory, and Prodi became prime minister. In 1998, Massimo D'Alema, leader of the Democrats of the Left (Democratici di Sinistra; DS), as the PDS had renamed itself, replaced Prodi. D'Alema's government was unpopular, and when the AN, LN, and Forza Italia formed a new alliance that crushed the Olive Tree in the 2000 regional elections, D'Alema resigned. Giuliano Amato assumed leadership of a weakened centre-left government, with the right well-poised for new elections in 2001. The right controlled all of the regions of the north, and there was growing pressure at the beginning of the 21st century for a redistribution of power from the central government to the regions.

Italy left the 20th century as a far richer and more developed country than a century earlier. Many problems remained, however, including continuing political instability and corruption, the historic but persistent economic and cultural divisions between the north and the south, and the new challenges of immigration and European unification. These challenges dominated Italy's political and economic agenda as it began a new century. (M.Cl./J.Fo./Ed.)

For later developments in the political history of Italy, see the BRITANNICA BOOK OF THE YEAR.

BIBLIOGRAPHY

Physical and human geography. *General works:* Comprehensive descriptions of many geographic aspects of the country are found in J.P. COLE, *Italy* (1964); DONALD SMITH WALKER,

A Geography of Italy, 2nd ed. (1967); PETER NICHOLS, *Italia, Italia* (1973); MARIO PINNA and DOMENICO RUOCCO (eds.), *Italy: A Geographical Survey* (1980); and JACQUES BÉTHEMONT and JEAN PELLETIER, *Italy: A Geographical Introduction* (1983; originally published in French, 1979). RINN S. SHINN (ed.), *Italy: A Country Study*, 2nd ed. (1987), is a comprehensive survey of both the geography and history of the country.

The people: National, social, and demographic characteristics of the country, with a look at regional differences, are discussed in DAVID WILLEY, *Italians* (1984); LUIGI BARZINI, *The Italians* (1964, reprinted 1977); WILLIAM MURRAY, *The Last Italian: Portrait of a People* (1991); DAVID I. KERTZER and RICHARD P. SALLER (eds.), *The Family in Italy from Antiquity to the Present* (1991); RUDOLPH M. BELL, *Fate and Honor, Family and Village: Demographic and Cultural Change in Rural Italy Since 1800* (1979); ROBERT E. DICKINSON, *The Population Problem in Southern Italy: An Essay in Social Geography* (1955, reprinted 1976); and ANN CORNELISEN, *Strangers and Pilgrims: The Last Italian Migration* (also published as *Flight from Torregreca: Strangers and Pilgrims*, 1980), which introduces a discussion of economic issues.

The economy: Broad surveys are offered in SHEPARD B. CLOUGH, *The Economic History of Modern Italy* (1964); PETER GROENEWEGER and JOSEPH HALEVI (eds.), *Italian Economics Past and Present* (1983); RUSSELL KING, *Italy* (1987); KEVIN ALLEN and ANDREW STEVENSON, *An Introduction to the Italian Economy* (1974); GISÈLE PODBIELSKI, *Italy: Development and Crisis in the Post-War Economy* (1974); VERA LUTZ, *Italy: A Study in Economic Development* (1962, reprinted 1975); JOHN EARLE, *Italy in the 1970s* (1975); and DONALD C. TEMPLEMAN, *The Italian Economy* (1981). Special studies include RUSSELL KING, *The Industrial Geography of Italy* (1985); and EDWARD GOODMAN, JULIA BAMFORD, and PETER SAYNOR (eds.), *Small Firms and Industrial Districts in Italy* (1989), both of these analyzing the location of industries; RAFFAELLA Y. NANETTI, *Growth and Territorial Policies: The Italian Model of Social Capitalism* (1988), focusing on regional disparities and the role of local government in economic development; ALAN B. MOUNTJOY, *The Mezzogiorno*, 2nd ed. (1982), briefly reviewing economic conditions in southern Italy; RUSSELL KING, *Land Reform, the Italian Experience* (1973); PINO ARLACCHI, *Mafia, Peasants, and Great Estates: Society in Traditional Calabria* (1983; originally published in Italian, 1980), and *Mafia Business: The Mafia Ethics and the Spirit of Capitalism* (1986; originally published in Italian, 1983); EDITH KURZWEIL, *Italian Entrepreneurs: Rearguard of Progress* (1983); and PETER LANGE and MARINO REGINI (eds.), *State, Market, and Social Regulation: New Perspectives on Italy* (1989; originally published in Italian, 1987). (R.L.K.)

Administration and social conditions: Comprehensive historical introductions to political institutions and processes are presented in NORMAN KOGAN, *A Political History of Italy: The Postwar Years*, new ed. (1983); PAUL GINSBORG, *A History of Contemporary Italy: Society and Politics, 1943–1988* (1990); FREDERIC SPOTTS and THEODOR WIESER, *Italy: A Difficult Democracy* (1986); and JOSEPH LAPALOMBARA, *Democracy, Italian Style* (1987). Particular aspects of the system are discussed in FRANCO DAMASO MARENGO, *Rules of the Italian Political Game* (1981); and JOHN FRASER, *Italy: Society in Crisis, Society in Transformation* (1981). Useful insights into the machinations of the power system are provided in ALAN FRIEDMAN, *Agnelli and the Network of Italian Power* (1988). ROBERT C. FRIED, *The Italian Prefects: A Study in Administrative Politics* (1963), examines local governmental structures. The dynamics of various political forces are the subject of PAOLO FARNETI, *The Italian Party System, 1945–1980* (1985); GEOFFREY PRIDHAM, *The Nature of the Italian Party System* (1981); ALAN S. ZUCKERMAN, *The Politics of Faction: Christian Democratic Rule in Italy* (1979); DONALD SASSOON, *The Strategy of the Italian Communist Party: From the Resistance to the Historic Compromise* (1981); GRANT AMYOT, *The Italian Communist Party: The Crisis of the Popular Front Strategy* (1981); SIDNEY TARROW, *Peasant Communism in Southern Italy* (1967); and JUDITH CHUBB, *Patronage, Power, and Poverty in Southern Italy: A Tale of Two Cities* (1982). A full analysis of the political, social, and economic crisis of the 1970s and the radical transformations that resulted from it is found in SIDNEY TARROW, *Democracy and Disorder: Protest and Politics in Italy, 1965–1975* (1989). The beginning of Italy's European integration is treated in F. ROY WILLIAMS, *Italy Chooses Europe* (1971). (G.Di P.)

Cultural life: ANNA LAURA LEPSCHY and GIULIO LEPSCHY, *The Italian Language Today*, 2nd ed. (1988), offers a survey of modern Italian and its dialects. A well-illustrated discussion of popular culture and social customs, with a look at rustic decoration and ornament, is found in CATHERINE SABINO, *Italian Country* (1988). DAVID FORGACS, *Italian Culture in the Industrial Era, 1880–1980: Cultural Industries, Politics, and the Public* (1990), is a history of popular culture and politico-cultural dynamics. JOHN JULIUS NORWICH (ed.), *The Italians: History, Art, and the Genius of a People* (also published as *The Italian World*, 1983), surveys developments from Roman times to the 20th century. JAMES HALL, *A History of Ideas and Images in Italian Art* (1983), traces the inspiration behind Italian art from Etruscan times. Other facets of cultural life are discussed in DOREEN YARWOOD, *The Architecture of Italy* (1970); J.H. WHITFIELD, *A Short History of Italian Literature*, 2nd ed. (1980); MICHAEL CAESAR and PETER HAINSWORTH (eds.), *Writers & Society in Contemporary Italy* (1984); ROBIN BUSS, *Italian Films* (1989); and PETER BONDANELLA, *Italian Cinema: From Neorealism to the Present*, new ed. (1990). (R.L.K./M.F.K.)

History. Extensive coverage of Italian history is provided in RUGGIERO ROMANO and CORRADO VIVANTI (eds.), *Storia d'Italia*, 6 vol. in 11 (1972–76), and *Storia d'Italia: Annali*, 9 vol. in 10 (1978–86), both published by Einaudi and divided by topics and periods; and *Storia d'Italia* (1979–), published by UTET and divided by political states. A comprehensive English-language survey is REINHOLD SCHUMANN, *Italy in the Last Fifteen Hundred Years: A Concise History*, 2nd ed. (1992).

Italy in the early Middle Ages: Comprehensive discussions of medieval Italy are provided in the first three volumes of *The Cambridge Medieval History*: vol. 1, *The Christian Roman Empire and the Foundation of the Teutonic Kingdoms*, 2nd ed. (1924, reprinted 1967), vol. 2, *The Rise of the Saracens and the Foundation of the Western Empire* (1913, reprinted 1967), and vol. 3, *Germany and the Western Empire* (1922, reprinted 1964). GIOVANNI TABACCO, *The Struggle for Power in Medieval Italy: Structures of Political Rule* (1989; originally published in Italian, 1979), is a major survey of sociopolitical history. Also of interest is CHRIS WICKHAM, *Early Medieval Italy: Central Power and Local Society, 400–1000* (1981).

Specific political and social topics of early periods are studied in A.H.M. JONES, *The Later Roman Empire, 284–602: A Social, Economic, and Administrative Survey*, 2 vol. (1964, reprinted 1986); T.S. BROWN, *Gentlemen and Officers: Imperial Administration and Aristocratic Power in Byzantine Italy, A.D. 554–800* (1984); THOMAS F.X. NOBLE, *The Republic of St. Peter: The Birth of the Papal State, 680–825* (1984); GUGLIELMO CAVALLO et al., *I bizantini in Italia* (1982); BARBARA M. KREUTZ, *Before the Normans: Southern Italy in the Ninth and Tenth Centuries* (1991); VITO FUMAGALLI, *Terra e società nell'Italia padana: i secoli IX e X* (1976); and several articles in the collective work *Structures féodales et féodalisme dans l'Occident méditerranéen: Xe–XIIIe siècles* (1980).

Important regional studies include CHRIS WICKHAM, *The Mountains and the City: The Tuscan Appennines in the Early Middle Ages* (1988); and PAOLO DELOGU, *Mito di una città meridionale: Salerno, secoli VIII–XI* (1977). Socioeconomic analyses include GINO LUZZATTO, *An Economic History of Italy: From the Fall of the Roman Empire to the Beginning of the Sixteenth Century* (1961; originally published in Italian, 1949); BRYAN WARD-PERKINS, *From Classical Antiquity to the Middle Ages: Urban Public Building in Northern and Central Italy, AD 300–850* (1984); and appropriate articles in RICHARD HODGES and BRIAN HOBLEY (eds.), *The Rebirth of Towns in the West, AD 700–1050* (1988). The art of this period is examined by RICHARD KRAUTHEIMER, *Rome, Profile of a City, 312–1308* (1980). (C.J.W.)

The High Middle Ages, 962–1300: Works embracing the whole of the period include *The Cambridge Medieval History*, vol. 5, *Contest of Empire and Papacy* (1926, reprinted 1968); GIORGIO FALCO, *The Holy Roman Republic: A Historic Profile of the Middle Ages*, 2nd ed. (1964, reprinted 1980; originally published in Italian, 1942); J.K. HYDE, *Society and Politics in Medieval Italy: The Evolution of the Civil Life, 1000–1350* (1973); and CHIARA FRUGONI, *A Distant City: Images of Urban Experience in the Medieval World* (1991; originally published in Italian, 1983). The reform era is studied in UTA-RENATE BLUMENTHAL, *The Investiture Controversy: Church and Monarchy from the Ninth to the Twelfth Century* (1988; originally published in German, 1982); and OVIDIO CAPITANI, *L'Italia medievale nei secoli di trapasso: la riforma della chiesa, 1012–1122* (1984). Politics of the 13th and 14th centuries are discussed in JOHN LARNER, *Italy in the Age of Dante and Petrarch, 1216–1380* (1980); and in the work in English by Tabacco, cited in the section above. The role of medieval Italy in Mediterranean commerce is examined in DAVID HERLIHY, ROBERT S. LOPEZ, and VSEVOLOD SLESSAREV (eds.), *Economy, Society, and Government in Medieval Italy* (1969); DAVID ABULAFIA, *Italy, Sicily, and the Mediterranean, 1100–1400* (1987); and GERALD W. DAY, *Genoa's Response to Byzantium, 1155–1204: Commercial Expansion and Factionalism in a Medieval City* (1988).

Examinations of social and cultural life in the 12th and 13th centuries include DAVID ABULAFIA, *The Two Italies: Economic Relations Between the Norman Kingdom of Sicily and the*

Northern Communes (1977); KARL BOSL, *Gesellschaftsgeschichte Italiens im Mittelalter* (1982); CHARLES T. DAVIS, *Dante's Italy, and Other Essays* (1984); JACQUES HEERS, *Parties and Political Life in the Medieval West,* trans. from French (1977); DAVID HERLIHY, *Medieval and Renaissance Pistoia: The Social History of an Italian Town, 1200–1430* (1967); PHILIP JONES, *Economia e società nell'Italia medievale* (1980); LESTER K. LITTLE, *Liberty, Charity, Fraternity: Lay Religious Confraternities at Bergamo in the Age of the Commune* (1988); LAURO MARTINES (ed.), *Violence and Civil Disorder in Italian Cities, 1200–1500* (1972); JAMES M. POWELL, *Albertanus of Brescia: The Pursuit of Happiness in the Early Thirteenth Century* (1992); and ALBERT RABIL, JR. (ed.), *Renaissance Humanism: Foundations, Forms, and Legacy,* vol. 1, *Humanism in Italy* (1988).

Broad comparative analyses of the communes are presented in GIAN PIERO BOGNETTI, *Studi sulle origini del comune rurale* (1978); RENATO BORDONE, *La società urbana nell'Italia comunale, secoli XI–XIV* (1984); RENATO BORDONE and JÖRG JARNUT (eds.), *L'evoluzione delle città italiane nell'XI secolo* (1988); LAURO MARTINES, *Power and Imagination: City-States in Renaissance Italy* (1979, reissued 1988); FERDINAND OPLL, *Stadt und Reich im 12. Jahrhundert, 1125–1190* (1986); YVES RENOUARD, *Les Villes d'Italie, de la fin du X^e siècle au début du XIV^e siècle,* new ed. edited by PHILIPPE BRAUNSTEIN, 2 vol. (1969); and DANIEL WALEY, *The Italian City-Republics,* 3rd ed. (1988).

Histories of individual city-states include MARIO DEL TREPPO and ALFONSO LEONE, *Amalfi medioevale* (1977); GEORGE W. DAMERON, *Episcopal Power and Florentine Society, 1000–1320* (1991); DUANE J. OSHEIM, *An Italian Lordship: The Bishopric of Lucca in the Late Middle Ages* (1977); CHRISTINE E. MEEK, *The Commune of Lucca Under Pisan Rule, 1342–1369* (1980); Violante's work on Milan, cited above among other regional studies; DANIEL WALEY, *Medieval Orvieto* (1952); ROBERT BRENTANO, *Rome Before Avignon: A Social History of Thirteenth-Century Rome* (1974, reissued 1991); FREDERICK C. LANE, *Venice, a Maritime Republic* (1973); and ALBERTO BROGLIO and LELLIA CRACCO (eds.), *Storia di Vicenza,* 4 vol. (1987–1990).

The Papal State is addressed in PETER PARTNER, *The Lands of St. Peter: The Papal State in the Middle Ages and the Early Renaissance* (1972); and DANIEL WALEY, *The Papal State in the Thirteenth Century* (1961). The history of the kingdom of Sicily is explored in H.E.J. COWDREY, *The Age of Abbot Desiderius: Montecassino, the Papacy, and the Normans in the Eleventh and Early Twelfth Centuries* (1983); JOSEF DEÉR, *Papsttum und Normannen: Untersuchungen zu ihren lehnsrechtlichen und kirchenpolitischen Beziehungen* (1972); PIETRO DE LEO, *Mezzogiorno medioevale: istituzioni, società, mentalità* (1984); DAVID C. DOUGLAS, *The Norman Achievement, 1050–1100* (1969), and *The Norman Fate, 1100–1154* (1976); NORMAN HOUSLEY, *The Italian Crusades: The Papal-Angevin Alliance and the Crusades Against Christian Lay Powers, 1254–1343* (1982); DONALD MATTHEW, *The Norman Kingdom of Sicily* (1992); L.-R. MÉNAGER, *Hommes et institutions de l'Italie normande* (1981); RAFFAELLO MORGHEN, *L'età degli svevi in Italia* (1974); *The Normans in Sicily and Southern Italy* (1977), a collection of three lectures; JOHN JULIUS NORWICH, *The Other Conquest* (also published as *The Normans in the South, 1016–1130,* 1967), and *The Kingdom in the Sun, 1130–1194* (1970); STEVEN RUNCIMAN, *The Sicilian Vespers: A History of the Mediterranean World in the Later Thirteenth Century* (1958, reprinted 1982); SALVATORE TRAMONTANA, *Mezzogiorno normanno e svevo* (1972); and HELENE WIERUSZOWSKI, *Politics and Culture in Medieval Spain and Italy* (1971). (J.M.Po.)

Italy in the 14th and 15th centuries: General outlines of the period, together with references to important works in Italian and other languages, are to be found in the work by Larner cited in the section above; and in DENYS HAY and JOHN LAW, *Italy in the Age of the Renaissance, 1380–1530* (1989). Significant general aspects are covered in several works cited above: the book by Waley on the city-republics; and the work by Martines and the one edited by him; and in MICHAEL MALLET, *Mercenaries and Their Masters: Warfare in Renaissance Italy* (1974). An excellent work of reference is J.R. HALE (ed.), *A Concise Encyclopaedia of the Italian Renaissance* (1981).

Works that consider the social background of high culture are JOHN LARNER, *Culture and Society in Italy, 1290–1420* (1971); PETER BURKE, *Tradition and Innovation in Renaissance Italy: A Sociological Approach,* rev. ed. (1974); and GEORGE HOLMES, *Florence, Rome, and the Origins of the Renaissance* (1986), and *The Florentine Enlightenment, 1400–50* (1969, reissued 1992). Influential discussions of the subject are DENYS HAY, *The Italian Renaissance in Its Historical Background,* 2nd ed. (1977); HANS BARON, *The Crisis of the Early Italian Renaissance: Civic Humanism and Republican Liberty in an Age of Classicism and Tyranny,* rev. ed. (1966); and PAUL OSKAR KRISTELLER, *Renaissance Thought: The Classic, Scholastic, and Humanistic Strains* (1961, reprinted 1980), and *Renaissance Thought II: Papers on Humanism and the Arts* (1961, reissued as *Renaissance Thought and the Arts,* 1980). The artist in society is discussed by D.S. CHAMBERS (compiler and trans.), *Patrons and Artists in the Italian Renaissance* (1970); and BRUCE COLE, *The Renaissance Artist at Work: From Pisano to Titian* (1983).

Most writings in English have concentrated on the republics. Valuable works on the great Tuscan city of Florence are GENE BRUCKER, *Florentine Politics and Society, 1343–1378* (1962) and *Renaissance Florence* (1969, reprinted 1983); RICHARD C. TREXLER, *Public Life in Renaissance Florence* (1980, reissued 1991); and DAVID HERLIHY and CHRISTIANE KLAPISCH-ZUBER, *Tuscans and Their Families: A Study of the Florentine Catasto of 1427* (1985; originally published in French, 1978); and, on the great ruling family, NICOLAI RUBINSTEIN, *The Government of Florence Under the Medici (1434–1494)* (1966); and J.R. HALE, *Florence and the Medici: The Pattern of Control* (1977, reissued 1983). An important treatment of a much-discussed character is DONALD WEINSTEIN, *Savonorola and Florence: Prophecy and Patriotism in the Renaissance* (1970).

D.S. CHAMBERS, *The Imperial Age of Venice, 1380–1580* (1970); and the work by Lane on Venice, cited above in the section on histories of the individual city-states, are two excellent general summaries on the Serenissima. Interesting contributions on the same topic are in OLIVER LOGAN, *Culture and Society in Venice, 1470–1790: The Renaissance and Its Heritage* (1972); J.R. HALE (ed.), *Renaissance Venice* (1973); ROBERT FINLAY, *Politics in Renaissance Venice* (1980); and EDWARD MUIR, *Civic Ritual in Renaissance Venice* (1981).

The Papal State is treated in the work by Partner cited in the section above; and studies of some of its *signori* include CECILIA M. ADY, *The Bentivoglio of Bologna: A Study in Despotism* (1937, reprinted 1969); JOHN LARNER, *The Lords of Romagna: Romagnol Society and the Origins of the Signorie* (1965); WERNER L. GUNDERSHEIMER, *Ferrara: The Style of a Renaissance Despotism* (1973); TREVOR DEAN, *Land and Power in Late Medieval Ferrara: The Rule of the Este, 1350–1450* (1988); and P.J. JONES, *The Malatesta of Rimini and the Papal State: A Political History* (1974). Other signorial regimes have been considered in D.M. BUENO DE MESQUITA, *Giangaleazzo Visconti, Duke of Milan (1351–1402): A Study in the Political Career of an Italian Despot* (1941); J.K. HYDE, *Padua in the Age of Dante* (1966); and LOUIS GREEN, *Castruccio Castracani: A Study on the Origins and Character of a Fourteenth-Century Italian Despotism* (1986).

Works on the southern kingdoms include DENIS MACK SMITH, *A History of Sicily,* vol. 1, *Medieval Sicily, 800–1713* (1968, reprinted 1988); BENEDETTO CROCE, *History of the Kingdom of Naples* (1970; originally published in Italian, 1925); and ALAN RYDER, *The Kingdom of Naples Under Alfonso the Magnanimous: The Making of a Modern State* (1976). (Jo.L.)

Early modern Italy (16th–18th centuries): Two magisterial works frame this period in its European perspective: FERNAND BRAUDEL, *The Mediterranean and the Mediterranean World in the Age of Philip II,* 2 vol. (1972–73; originally published in French, 2nd rev. and expanded ed., 1966); and FRANCO VENTURI, *Settecento riformatore* (1969–), with English translations of 2 vol., *The End of the Old Regime in Europe, 1768–1776: The First Crisis* (1989) and *The End of the Old Regime in Europe, 1776–1789,* 2 parts (1990–91).

General surveys of the period include the work by Hay and Law cited in the section above; ERIC COCHRANE, *Italy, 1530–1630,* ed. by JULIUS KIRSHNER (1988); and DINO CARPANETTO and GIUSEPPE RICUPERATI, *Italy in the Age of Reason, 1685–1789* (1987); and STUART WOOLF, *A History of Italy, 1700–1860: The Social Constraints of Political Change* (1979, reissued 1991).

Studies of the individual states include, on Savoy, GEOFFREY SYMCOX, *Victor Amadeus II: Absolutism in the Savoyard State, 1675–1730* (1983); on Venice, the work by Lane cited above in the section on histories of the individual city-states; on Florence, ERIC COCHRANE, *Florence in the Forgotten Centuries, 1527–1800* (1973); on Milan, DOMENICO SELLA, *Crisis and Continuity: The Economy of Spanish Lombardy in the Seventeenth Century* (1979); on Rome, PETER PARTNER, *Renaissance Rome, 1500–1559: A Portrait of a Society* (1976); and HANNS GROSS, *Rome in the Age of Enlightenment: The Post-Tridentine Syndrome and the Ancien Regime* (1990); on Naples, ANTONIO CALABRIA and JOHN A. MARINO (eds. and trans.), *Good Government in Spanish Naples,* trans. from Italian (1990); and ROSARIO VILLARI, *The Revolt of Naples* (1993; originally published in Italian, 1967); and, on Sicily, H.G. KOENIGSBERGER, *The Practice of Empire* (1969). Of special interest is CARLO GINZBURG, *The Cheese and the Worms: The Cosmos of a Sixteenth-Century Miller* (1980, reissued 1992; originally published in Italian, 1976). (J.A.M.)

Revolution, restoration, and unification: Works covering this period include GEORGE MARTIN, *The Red Shirt and the Cross of Savoy: The Story of Italy's Risorgimento, 1748–1871* (1969);

the work by Woolf on political change cited in the previous section; JOHN A. DAVIS and PAUL GINSBORG (eds.), *Society and Politics in the Age of the Risorgimento* (1991), a collection of essays; FRANK J. COPPA, *The Origins of the Italian Wars of Independence* (1992); RAYMOND GREW, *A Sterner Plan for Italian Unity* (1963); CLARA M. LOVETT, *The Democratic Movement in Italy, 1830–1876* (1982), and *Giuseppe Ferrari and the Italian Revolution* (1979); DENIS MACK SMITH, *Cavour and Garibaldi, 1860: A Study in Political Conflict* (1954, reissued 1985); and PAUL GINSBORG, *Daniele Manin and the Venetian Revolution of 1848–49* (1979). Entries on major events and figures may be found in FRANK J. COPPA (ed.), *Dictionary of Modern Italian History* (1985). (C.M.Lo.)

Italy since 1870: The fullest one-volume studies in English are DENIS MACK SMITH, *Italy: A Modern History*, new ed. rev. and enlarged (1969); and MARTIN CLARK, *Modern Italy, 1871–1982* (1984). EDWARD R. TANNENBAUM and EMILIANA P. NOETHER (eds.), *Modern Italy: A Topical History Since 1861* (1974), covers particular aspects; as do DENIS MACK SMITH, *Italy and Its Monarchy* (1989); and the work by Clough cited above in the section on the economy. CHRISTOPHER SETON-WATSON, *Italy from Liberalism to Fascism, 1870–1925* (1967, reprinted 1981); and C.J. LOWE and F. MARZARI, *Italian Foreign Policy, 1870–1940* (1975), are excellent on the first half of the period. The best work in Italian is that by GIORGIO CANDELORO, *Storia dell'Italia moderna*, vol. 6–11 (1970–86), which covers the years from 1871 to 1948.

The pre-1914 period is addressed in RAFFAELE ROMANELLI, *L'Italia liberale (1861–1900)* (1979), and *Il comando impossibile* (1988). Foreign policy before World War I is the topic of R.J.B. BOSWORTH, *Italy, the Least of the Great Powers* (1979), and *Italy and the Approach of the First World War* (1983). The early socialist movement is discussed well in RICHARD HOSTETTER, *The Italian Socialist Movement*, vol. 1, *Origins (1860–1882)* (1958); and LOUISE A. TILLY, *Politics and Class in Milan, 1881–1901* (1992). The army is treated in JOHN GOOCH, *Army, State, and Society in Italy, 1870–1915* (1989). An excellent overview of economic history is given in GIANNI TONIOLO, *An Economic History of Liberal Italy, 1850–1918* (1990; originally published in Italian, 1988); and GIORGIO FUÀ (ed.), *Lo sviluppo economico in Italia*, 3 vol. (1969–81).

EDWARD R. TANNENBAUM, *The Fascist Experience* (1972, also published as *Fascism in Italy*, 1973); and LUISA PASSERINI, *Fascism in Popular Memory* (1987; originally published in Italian, 1984), are interesting social histories. DENIS MACK SMITH, *Mussolini* (1981), is indispensable; more detail is available in RENZO DE FELICE, *Mussolini* (1965–), in Italian. The rise of fascism in particular is examined by ADRIAN LYTTELTON, *The Seizure of Power*, 2nd ed. (1987); and ALEXANDER J. DE GRAND, *The Italian Nationalist Association and the Rise of Fascism in Italy* (1978). Good local studies are PAUL CORNER, *Fascism in Ferrara, 1915–1925* (1975); ANTHONY L. CARDOZA, *Agrarian Elites and Italian Fascism: The Province of Bologna, 1901–1926* (1982); and FRANK M. SNOWDEN, *The Fascist Revolution in Tuscany, 1919–1922* (1989).

The diversity of fascist ideas is revealed in works by ADRIAN LYTTELTON (ed.), *Italian Fascisms from Pareto to Gentile* (1973); RENZO DE FELICE, *Interpretations of Fascism* (1977; originally published in Italian, 1969); and DAVID D. ROBERTS, *The Syndicalist Tradition in Italian Fascism* (1979); and in biographies of leading fascists—*e.g.*, CLAUDIO G. SEGRÈ, *Italo Balbo* (1987). The impact of fascism is discussed by VICTORIA DE GRAZIA, *The Culture of Consent* (1981), and *How Fascism Ruled Women* (1992); and DOUG THOMPSON, *State Control in Fascist Italy* (1991).

Church-state relations are examined in D.A. BINCHY, *Church and State in Fascist Italy* (1941, reprinted 1970); and JOHN F. POLLARD, *The Vatican and Italian Fascism, 1929–32* (1985). Anti-Semitism is treated in MEIR MICHAELIS, *Mussolini and the Jews* (1978). Fascist foreign policy is the subject of DENIS MACK SMITH, *Mussolini's Roman Empire* (1976).

CHARLES F. DELZELL, *Mussolini's Enemies: The Italian Anti-Fascist Resistance* (1961), is still useful on the Resistance period. DAVID W. ELLWOOD, *Italy, 1943–1945* (1985), is strong on diplomatic history. The best history of Italian communism is PAOLO SPRIANO, *Storia del Partito Comunista Italiano*, 5 vol. (1967–75). An English work on this topic is ALEXANDER J. DE GRAND, *The Italian Left in the Twentieth Century* (1989).

Chapters in S.J. WOOLF (ed.), *The Rebirth of Italy, 1943–50* (1972), discuss postwar reconstruction; as does JOHN LAMBERTON HARPER, *America and the Reconstruction of Italy, 1945–1948* (1986). PETER LANGE and SIDNEY TARROW (eds.), *Italy in Transition* (1980), is strong on social changes.

The best accounts of the postwar political system are in FREDERIC SPOTTS and THEODOR WIESER, *Italy: A Difficult Democracy* (1986); GIORGIO GALLI and ALFONSO PRANDI, *Patterns of Political Participation in Italy* (1970); and DAVID HINE, *Governing Italy* (1993). Studies of the economy include the work by Allen and Stevenson cited above in the section on the economy; PETER LANGE and MARINO REGINI, *State, Market, and Social Regulation* (1989); and JUDITH CHUBB, *Patronage, Power, and Poverty in Southern Italy* (1982). The classic text on regionalism is now ROBERT D. PUTNAM, ROBERT LEONARDI, and RAFAELLA Y. NANETTI, *Making Democracy Work* (1993).

On the Mafia and organized crime, particularly noteworthy are MICHELE PANTALEONE, *The Mafia and Politics* (1966; originally published in Italian, 1962); HENNER HESS, *Mafia and Mafiosi* (1973; originally published in German, 1970); ANTON BLOK, *The Mafia of a Sicilian Village, 1860–1960* (1974, reissued 1988); GIOVANNI FALCONE and MARCELLE PADOVANI, *Men of Honour* (1992; originally published in Italian, 1991); and the work by Arlacchi cited above in the section on the economy. The best studies of terrorism in the 1970s are DAVID MOSS, *The Politics of Left-Wing Violence in Italy, 1969–85* (1989); and RAIMONDO CATANZARO (ed.), *The Red Brigades and Left-Wing Terrorism in Italy* (1991). (M.Cl.)

Jainism

Jainism—along with Hinduism and Buddhism—is one of the three most ancient of India's religious traditions still in existence. Its name derives from the Sanskrit verb root *ji,* "to conquer." The name refers to the ascetic battle that the Jaina monks must fight against the passions and bodily senses in order to gain omniscience and the complete purity of soul that represents the highest religious goal in the Jaina system. The monk-ascetic who achieves this omniscience and purity is called a Jina (literally, "Conqueror," or "Victor"), and adherents to the tradition are called Jainas, or Jains. Although Jainism has a much smaller number of adherents than do Hinduism and Sikhism, its influence on India's culture has been considerable, including significant contributions in philosophy and logic, art and architecture, grammar, mathematics, astronomy and astrology, and literature.

Jainism has largely been confined to India, although the migration of Indians to other, predominantly English-speaking countries has spread its practice to many Commonwealth nations and to the United States. Its continuous existence in India for some 2,500 years is in sharp contrast to Buddhism, which is widespread in Asia but no longer widely practiced in the land of its origin. This gives Jainism a unique status as the only Sanskritic non-Hindu religious tradition to have survived in India to the present.

For coverage of related topics in the *Macropædia* and *Micropædia,* see the *Propædia,* section 823.

This article is divided into the following sections:

HISTORY

Early history (6th century BC–c. 5th century AD). Jaina history began in the 6th century BC with Vardhamāna, who is known as Mahāvīra ("Great Hero"). Mahāvīra was the 24th and last Tirthankara (literally, "Ford-maker") of the current age (*kalpa*) of the world. (Tirthankaras, also called Jinas, are revealers of the Jaina religious path [*dharma*] who have crossed over life's stream of rebirths and have set the example that all Jainas must follow.) Mahāvīra was a contemporary of Siddhārtha Gautama (the Buddha) and was born in the same area, the lower Gangetic Plain. Although Mahāvīra was a historical figure, all of the accounts of his life are legendary and serve the ritual life of the Jaina community better than they do the historian. However, a little of the historical circumstances of Mahāvīra and the early Jaina community can be pieced together from a variety of sources.

Mahāvīra

The 6th century BC was a period of intense religious activity in the lower Gangetic Plain. In addition to Buddhism, the Ājīvika sect, founded by Gośāla Maskarīputra, appeared; and at about this time, probably in the same region, the two great "forest" Upanishad texts of early Hinduism, the *Brihadāranyaka* and the *Chāndogya,* came into existence. The prevailing ethos common to all these religious perspectives was asceticism, which stood in contrast to the ritualistic Brahmanic schools associated with the earliest period of classical Hinduism.

Mahāvīra, like the Buddha, was the son of a chieftain of the Kshatriya (military or ruling) class. At age 30 he renounced his princely status to take up the ascetic life. It is likely that he pursued the discipline of a preestablished ascetic tradition and had a reforming influence on it. His acknowledged status as the 24th Tirthankara (or Jina) means that Jainas perceive him as the last revealer in this cosmic age of the Jaina *dharma.* Mahāvīra had 11 disciples (called *gaṇadharas*), all of whom were Brahman converts to Jainism; all founded monastic lineages, but only two—Indrabhūti Gautama and Sudharman, the disciples who survived Mahāvīra—served as the points of origin for the historical Jaina monastic community.

The community appears to have grown quickly—Jaina tradition states that it numbered 14,000 monks and 36,000 nuns at the time of Mahāvīra's death. From the beginning the community was subject to a number of schismatic movements. Jamāli, Mahāvīra's son-in-law, led the first of seven schisms that occurred during the Jina's lifetime. None of these had a significant effect on the Jaina community. The only schism to have a lasting effect was that between the Svetambaras (literally, "White-robed") and the Digambaras ("Sky-clad"; *i.e.,* naked); this division still exists. The major points of difference between the two concern the question of proper monastic attire and whether or not a soul can attain liberation from a female body (a possibility the Digambaras deny).

Each sect has its own account of how the schism arose. It appears to have begun as a physical split of the community during the 3rd century BC. According to Digambara tradition, Bhadrabāhu I (whom the Digambaras regard as their founder) foresaw a 12-year famine in the Mauryan kingdom of Candra Gupta and took half of the monastic community south with him to Śravaṇa Beḷgoḷa (near modern Hassan, in Karnātaka state). Digambara tradition also states that Candra Gupta accompanied Bhadrabāhu as his disciple. Svetambara tradition, however, states that Bhadrabāhu went to Nepal and that the Svetambara–Digambara split was led by a monk named Śivabhuti in the last half of the 1st century AD. All differences of doctrine and praxis between the two sects appear to have arisen from this geographical separation.

The four councils

These differences were formalized through a series of councils that met to preserve and codify the teachings of Mahāvīra in written form. It was felt that the teachings, preserved orally since his death, were in danger of being lost. Four councils were held between the 4th century BC and the 5th century AD. The last one, held at Valabhī in Saurāṣṭra (modern Gujarāt state) in either AD 453 or 456, codified the Svetambara canon that is still in use. The Digambara monastic community considered this redaction too corrupt to be normative, and the schism between the two communities became irrevocable.

During this period, Jainism spread from its place of origin westward to Ujjain, where it gained the patronage of Candra Gupta, the grandfather of Aśoka (the last great Mauryan emperor), and later Samprati, the grandson of Aśoka. Later, in the 1st century BC, a monk named Kālakācārya seems to have caused the overthrow of King Gardabhilla of Ujjain and his replacement with the Śāhi kings, who were probably of Scythian or Persian origin. By the time of the Gupta dynasty (AD 320–*c.* 600), Jainas were retaining the patronage of the Gupta emperors of Magadha, but they had become stronger in central and western India than in their homeland.

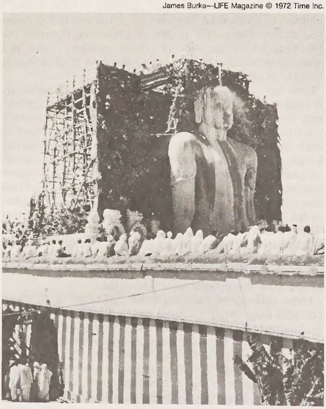
Early medieval developments (500–1100). The early medieval period was the time of Jainism's greatest flowering, particularly for the Digambara community in the south. The Digambaras gained the patronage of three major dynasties during these centuries—the Gaṅgas in Karnātaka (3rd–11th century); the Rāṣṭrakūṭas, whose kingdom was just north of the Gaṅga realm (8th–12th century); and the Hoysaḷas in Karnātaka (11th–14th century). Digambara monks are reputed to have engineered the succession of the Gaṅga and the Hoysaḷa dynasties, thus stabilizing uncertain political situations and guaranteeing Jaina political protection and support.

Digambara political activity

This involvement in politics on the part of the Digambaras allowed Jainism to prosper in Karnātaka and the Deccan. An abundance of epigraphical evidence details an elaborate patronage system through which kings, queens, state ministers, and military generals endowed the Jaina community with tax revenues and with direct grants for the construction and upkeep of temples. In addition, many of these political figures had Jaina monks as spiritual teachers and advisers. Two notable examples are Śāntala Devī, the queen of the Hoysaḷa king Viṣṇuvardhana, and the Gaṅga general Chāmuṇḍarāya, who in the 10th century oversaw the creation of a colossal statue of Bāhubali (locally called Gomateśvara; son of Ṛṣabha, the first Tirthankara) at Śravaṇa Beḷgoḷa.

During this period Digambara writers produced a large amount of philosophical treatises, commentaries, and poetry, which was written in Prakrit, Kannada, and Sanskrit. Much of this literary activity had royal patronage and participation. Noteworthy was the monk Jinasena, whose Sanskrit philosophical and poetic writing had the support of the Rāṣṭrakūṭa king, Amoghavarṣa I. Himself an author in Kannada and Sanskrit, Amoghavarṣa seems to have renounced his throne and become a disciple of Jinasena in the early 9th century. This privileged position allowed the Digambara Jainas to engage in sectarian debate from a position of strength. Inscriptions and epigraphs describe many of the most important monks of this period as victors over the Buddhists, Brahmans, Vaiṣṇavites, and Śaivites in philosophical and religious debate.

James Burke—LIFE Magazine © 1972 Time Inc.



Ceremony of anointing the colossal image of the Jaina saint Bāhubali (called locally Gommateśvara) at Śravaṇa Beḷgoḷa, Karnātaka state, India.

The Svetambaras seem to have been less flamboyantly embroiled in dynastic politics than their southern counterparts, though there is evidence of such activity in Gujarāt and Rājasthān that helped establish sympathetic kings in the 8th century (Vānarāja, 716–806) and the 12th century (Kumārapāla, whose reign ended with the Muslim invasions). Kumārapāla's accession was masterminded by the great Svetambara scholar and minister of state Hemacandra. The Svetambaras were no less productive in literary output than their Digambara counterparts at this period.

Beginning in the early centuries AD, the role of the Jaina layman was articulated with a detail and precision not seen up to that point. The process began for the Digambaras as early as the 2nd to 3rd century; with the Svetambaras it seems to have begun in the 5th to 6th century. The early medieval period was a time of particularly intense reflection for both groups on the role of the laity. A large Āvaśyaka literature, discussing the layman's religious behaviour and vows, poured forth from these beginnings and lasted until the 17th century. A formalized caste system appeared among the Jaina laity. This was depicted and given authority by Jinasena in his *Ādipurāṇa,* a hagiography of Ṛṣabha and his two sons Bāhubali and Bharata. It differed from the Hindu system in that the Kshatriyas were given a place of prominence over the Brahmans; in addition, the Jainas did not see the caste system as an inherent part of the structure of a created universe. There also were differences in the organization of the caste system between the Digambaras and the Svetambaras.

Late medieval–early modern developments (1100–1800). In the period of their greatest influence (6th–late 12th century), Jaina monks ceased being wandering ascetics and tended to become dwellers at temples or monastic residences, surrounded by the comforts that their calling demanded they forego. In addition, the Digambara monks' active involvement in dynastic politics undoubtedly earned them enemies. These two factors led to a decline of Jaina influence in ensuing centuries.

Effect of the Muslim invasion

The Svetambara community's eclipse was greatly accelerated by the successful invasion of Muslim forces into western and northern India in the 12th century. With this sudden shift of political control from indigenous to foreign hands, the Svetambara community concentrated on stabilizing itself in the new circumstances. At about this time, the monastic libraries were put underground in Rājasthān to keep the manuscripts from being destroyed and to preserve them better from the elements. There is evidence of Jaina laymen serving as ministers to Muslim rulers, which surely benefited the community.

Reform movements appeared within the community at various times, often stressing the inappropriateness of image worship, especially for monks. This was likely a response to strong Muslim religious values. The most successful of these reform movements was that of the mid-15th-century layman Lonkā Sāha, which led ultimately to the founding of the Sthānakavāsī sect in the 18th century.

By the advent of the Vijayanagar Empire in the 14th century, Digambara Jainism had lost all significant royal support and survived largely by keeping to itself. At this time elaborate temple rituals with Tantric overtones developed within the Digambara community. This, plus the lax attitudes of the administrators of Digambara temple complexes, helped fuel resentment both within and outside the community. This situation made the Digambaras susceptible to attacks by renascent Hindu devotional movements. These movements began in Tamil Nādu as early as the 6th century and in Karnātaka in the 12th century. One of the most vigorous of these Hindu movements was that of the Lingayats, or Vīraśaivas, which arose in full force in the 12th century in northern Karnātaka, a stronghold of Digambara Jainism. The Lingayats gained royal support, and many Jainas themselves converted to the Lingayat religion in ensuing centuries.

Digambara laity were among the foremost critics of their community's deteriorating situation. The most significant Digambara reform movement occurred in the late 16th century, led by a layman and poet named Bānārasīdās. This movement attacked the elaborateness of Digambara ritualism and the cavalier behaviour of its religious leaders.

Recent Jaina history. In modern times, Svetambara Jainism has maintained a more effective organization and has a larger monastic community than its Digambara counterpart. Both communities devote much energy to maintaining temples and publishing critical editions of their religious texts.

In addition, the Jainas stress publicly their deep and long-standing commitment to *ahiṃsā* ("nonviolence"). Notable in this connection is the friendship and exchange of letters between Mohandas Gandhi and the Svetambara layman Raychandrabhai Mehta. Gandhi considered his interactions with Mehta to be important in formulating his own ideas on the use of nonviolence as a political tactic.

Jainas have traditionally been professional and mercantile people. These trades have made them adaptable to other environments and societies besides those of India. Many Jainas have emigrated overseas, and this has had the result of increasing international awareness of Jainism. (G.R.S.)

IMPORTANT FIGURES OF JAINA LEGEND

Sixty-three significant figures form the focus of Jaina legend and story. The most important of these are the 24 Tirthankaras, perfected human beings who appear from time to time to preach and embody the Jaina religious path; they represent the highest religious attainment for the Jaina. The Tirthankaras, along with 12 *cakravartins* ("world conquerors"), nine *vāsudevas* (counterparts of Vāsudeva, the patronymic of Krishna), and nine *baladevas* (counterparts of Balarāma, the elder half-brother of Krishna), constitute a list of 54 *mahāpuruṣas* ("great souls"), to which were later added nine *prativāsudevas* (enemies of the *vāsudevas*). Other, more minor, figures include nine *nāradas* (counterparts of the deity Nārada, the messenger between gods and humans), 11 *rudras* (counterparts of the Vedic god Rudra, from whom Śiva is said to have evolved), and 24 *kāmadevas* (gods of love), all of which show Hindu influences. Bāhubali is said to be the first *kāmadeva*.

Subordinated to these figures are the gods, which are classified into four groups: the *bhavanavāsīs* (gods of the house), the *vyantaras* (intermediaries), the *jyotiṣkas* (luminaries), and the *vaimānikas* (astral gods). These, in turn, are divided into several subgroups. Other gods and goddesses also occur in various Jaina texts, such as the 64 *dikkumārīs* (maidens of the directions), who act as nurses to a newborn Tirthankara. Such deities played an important role in ancient Indian folk religion, and the Jainas, Buddhists, and Hindus all assimilated them into their pantheons and rituals.

DOCTRINES OF JAINISM

The Jaina's religious goal is the complete perfection and purification of the soul. This can occur only when the soul is in a state of eternal liberation from and nonattachment to corporeal bodies. Liberation of the soul is impeded by the accumulation of *karmans* (see below *Karman*), bits of material, generated by a person's actions, that bind themselves to the soul and consequently bind the soul to material bodies through many births; this has the effect of thwarting the full self-realization and freedom of the soul. To understand how the Jainas perceive and address this problem, it is first necessary to explain the Jaina conception of reality.

The notion of the wheel of time

Time and the universe. Time, according to the Jainas, is eternal and formless. It is conceived as a wheel with 12 spokes called *ārās* ("ages"), six making an ascending arc and six a descending one. In the ascending arc (*utsarpiṇī*), man progresses in knowledge, age, stature, and happiness, while in the descending arc (*avasarpiṇī*) he deteriorates. The two cycles joined together make one rotation of the wheel of time, which is called a *kalpa*.

The world is eternal and uncreated. Its constituent elements, the six substances (*dravyas*), are soul, matter, time, space, and the principles of motion and the arrest of motion. These are eternal and indestructible, but their conditions change constantly.

Jainas divide the inhabited universe into five parts. The lower world (*adholoka*) is subdivided into seven tiers, each one darker and more tortuous than the one above it. The middle world (*madhyaloka*) consists of numberless concentric continents separated by seas, the centre continent of which is called Jambudvīpa. Human beings occupy Jambudvīpa, the second continent, and half of the third; the focus of Jaina activity, however, is Jambudvīpa, the only continent on which it is possible for the soul to achieve liberation. The celestial world (*ūrdhvaloka*) consists of two categories of heaven: one for the souls of those who may or may not have entered the Jaina path and one for the souls of those who are far along on the path and are close to the time of their emancipation. At the apex of the occupied universe is the *siddhaśīlā*, the crescent-shaped abode of liberated souls (*siddhas*). Finally, there are some areas inhabited solely by *ekendriyas*, organisms that have only a single sense. Although *ekendriyas* permeate all parts of the occupied universe, there are places where they are the only living beings.

Jīva and ajīva. Jaina reality is constituted by *jīva* ("soul," or "living substance") and *ajīva* ("non-soul," or "inanimate substance"). *Ajīva* is divided into two categories: (1) non-sentient and material and (2) non-sentient and nonmaterial. All but *jīva* are without life.

Characteristics of *jīva*

The essential characteristics of *jīva* are consciousness (*cetanā*), bliss (*sukha*), and energy (*vīrya*). In its pure state, *jīva* possesses these qualities in infinite measure. The souls, infinite in number, are divisible in their embodied state into two main classes, immobile and mobile, according to the number of sense organs possessed by the body they inhabit. The first group consists of souls inhabiting immeasurably small particles of earth, water, fire, and air, plus the vegetable kingdom, which possess only the sense of touch. The second group comprises souls that inhabit bodies that have between two and five sense organs. The Jainas believe that the four elements (earth, water, fire, and air) also are animated by souls. Moreover, the universe is full of an infinite number of minute beings, *nigodas*, which are slowly evolving.

A *jīva* is formless and genderless and cannot be perceived by the senses. A soul is not all-pervasive but can, by contraction or expansion, occupy various amounts of space. Like the light of a lamp in a small or a large room, it can fill both the smaller and larger bodies it occupies. While the soul assumes the exact dimensions of the body it occupies, it is not identical with that body.

Matter (*pudgala*) has the characteristics of touch, taste, smell, and colour. Its essential characteristic is lack of consciousness. The smallest unit of matter is the atom (*paramāṇu*). Heat, light, and shade are forms of fine matter.

The non-sentient, nonmaterial substances are the principles of motion and its arrest, space, and time. They are always pure and are not subject to defilement. The principles of motion and its arrest permeate the universe; they do not exist independently but, rather, form a necessary precondition for any object's movement or coming to rest. Space is infinite, all-pervasive, and formless and provides accommodation for the entire universe. It is divided into occupied (*i.e.*, the universe) and unoccupied portions. Time is said to consist of innumerable eternal and indivisible particles of "non-corporeal substance" that never mix with one another but that fill the entire universe. Thus, the non-sentient, nonmaterial substances form the context within which occurs the drama of a *jīva*'s struggle to extricate itself from involvement with matter.

Karman. The fundamental tenet of Jaina doctrine is that all phenomena are linked together in a universal chain of cause and effect. Every event has a definite cause behind it. By nature each soul is pure, possessing infinite knowledge, bliss, and power; however, these faculties are restricted from beginning-less time by foreign matter coming in contact with the soul. Fine foreign matter producing the chain of cause and effect, of birth and death, is *karman*, a fine atomic substance and not a process as in Hinduism. To be free from the shackles of *karman*, a person must stop the influx of new *karmans* and eliminate the acquired ones.

Karmic particles are acquired as the result of intentional action tinged with passionate expression. Acquired *karmans* can be annihilated through a process called *nirjarā*, which consists of fasting, not eating certain kinds of food, control over taste, resorting to lonely places, mortifications of the body, atonement and expiation for sins, modesty, service, study, meditation, and renunciation of the ego. *Nirjarā* is, thus, the calculated cessation of passionate action.

Stages of the soul's growth

A soul passes through various stages of spiritual development before becoming free from all karmic bondages. These stages of development (*guṇasthānas*) involve progressive manifestations of the innate faculties of knowledge and power and are accompanied by decreasing sinfulness and increasing purity.

Jīvas become imprisoned in a succession of bodies owing to their connection with karmic matter. These embodied souls bear different colours or tints (*leśyā*), varying according to the merits or demerits of the particular being. This doctrine of *leśyā*, peculiar to Jainism, seems to have been borrowed from the Ājīvika doctrine of six classes of bodies, expounded by Gośāla Maskarīputra. The six *leśyās* in Jainism are, in the ascending order of man's spiritual progress, black, blue, gray, fiery red, lotus-pink (or yellow), and white.

Theories of knowledge as applied to liberation. In Jaina thought, four stages of perception—observation, will to recognize, determination, and impression—lead to a subjective cognition (*matijñāna*), the first of five kinds of knowledge (*jñāna*). The second kind of knowledge is *śrutajñāna*, derived from the scriptures and general information. Both of these are mediated cognition, based on external conditions perceived by the senses. There are three kinds of immediate knowledge—*avadhi* (supersensory perception), *manaḥparyāya* (reading the thoughts of others), and *kevala*, which is the stage of omniscience. *Kevala* is necessarily accompanied by freedom from karmic obstruction and by direct experience of the soul's pure form unblemished by its attachment to matter. Omniscience is the foremost attribute of a liberated *jīva*, the emblem of its purity; thus, a liberated soul, such as a Tirthankara, is called a *kevalin* ("possessor of omniscience").

According to Jainism, *yoga*, the ascetic physical and meditative discipline of the monk, is the means to the attainment of omniscience and thus to *mokṣa*, or liberation. *Yoga* is the cultivation of true knowledge of reality, faith in the teachings of the Tirthankaras, and pure conduct; it is, thus, intimately connected to the three jewels (*ratnatraya*) of right knowledge, right belief, and right conduct (respectively, *samyagjñāna*, *samyagdarśaṇa*, and *samyakcāritra*). (See INDIAN PHILOSOPHY.)

Jaina ethics. The *ratnatraya* constitute the basis of Jaina ethics. Right knowledge, faith, and conduct must be cultivated together; none of them can be achieved in the absence of the others. Right faith leads to calmness or tranquillity, detachment, kindness, and the renunciation of pride of birth, beauty of form, wealth, scholarship, prowess, and fame. Right faith leads to perfection only when followed by right conduct. Yet, there can be no virtuous conduct without right knowledge, which consists of clear distinction between the self and the nonself. Knowledge of scriptures is distinguished from inner knowledge. Knowledge without faith and conduct is futile. Without purification of mind, all austerities are mere bodily torture. Right conduct is thus spontaneous, not a forced mechanical quality. Attainment of right conduct is a gradual process, and a householder can observe only partial self-control; when he becomes a monk, he is further able to observe more comprehensive rules of conduct.

Concept of *ahiṃsā*

Two separate courses of conduct are laid down for the ascetics and the laity. In both cases the code of morals is based on the doctrine of *ahiṃsā*, or nonviolence. Since thought gives rise to action, violence in thought merely precedes violent behaviour. Violence in thought, then, is the greater and subtler form of violence, because it arises from ideas of attachment and aversion, grounded in passionate states, which result from negligence or lack of care in behaviour. Jainism enjoins avoidance of all forms of injury, whether committed by body, mind, or speech.

RITUAL PRACTICES AND RELIGIOUS AND SOCIAL INSTITUTIONS

The monks and their practices. Svetambaras acknowledge two classes of monks: *jinakalpins*, who wander naked and use the hollows of their palms as alms bowls; and *sthavirakalpins*, who retain minimal possessions such as a robe, an alms bowl, a whisk broom, and a *mukhavastrikā* (a piece of cloth held over the mouth to protect against the ingestion of small insects). A monk must obey the "great vows" (*mahāvratas*) to avoid injuring any life-form, lying, stealing, having sexual intercourse, or accepting personal possessions. To help him live out his vows, a monk's life is carefully regulated in all details by specific ordinances and by the oversight of his superiors. For example, to help him observe the vow of noninjury, a monk may not take meals after dark, since to do so would increase the possibility that he would harm any insects that might be attracted to the food. Monks are expected to suffer with equanimity such hardships as those imposed by the weather, geographic terrain, travel, or physical abuse. Exceptions are allowed in emergencies, since a monk who survives a calamity can purify himself by confession and by practicing even more rigorous austerities.

Among the Digambaras, a full-fledged monk remains naked, though there are lower-grade monks who wear a loincloth and keep with them one piece of cloth not more than one and one-half yards long. Digambara monks use a peacock-feather duster and water gourd, live apart from human habitations, and beg and eat only once a day, using the palm of one hand as an alms bowl.

Eight essentials noted for the conduct of monks include the three *guptis* (care in thought, speech, and action) and the five *samitis* (kinds of vigilance over conduct). The six *āvaśyakas*, or obligations, are equanimity; praise of the Tirthankaras (Jinas); obeisance to the Jinas, teachers, and scriptures; atonement; resolution to avoid sinful activities; and meditation.

The type of austerities in which a monk engages, the length of time he engages in them, and their severity are carefully regulated by his preceptor, who takes into account the monk's spiritual development, his capacity to withstand the austerities, and his ability to understand how they help further his spiritual progress at a given time.

The act of *sallekhanā*

The culmination of a monk's ascetic rigours is the act of *sallekhanā*, in which a monk lies on one side on a bed of thorny grass and ceases to move or take food. This act of ritual starvation is the monk's ultimate act of nonattachment, in which he lets go of the body for the sake of his soul. The ascetic's preparatory rigours, which point to and culminate in this act, generally take 30 years or more to perform. Although it is a tenet of Jaina doctrine that no one can achieve liberation in this corrupt time, it is thought that the act of *sallekhanā* nevertheless has value, because it can improve a soul's spiritual situation in the next birth.

Religious disciplines of the laity. The life of a lay votary is a preparatory stage to the rigours of ascetic life. The lay votary is enjoined to observe eight primary behavioral qualities (which vary but usually include the avoidance of meat, wine, honey, fruits, roots, and night eating) and 12 vows: five *aṇuvratas* ("little vows"), three *guṇavratas*, and four *śikṣāvratas*. The *aṇuvratas* are vows to abstain from gross violence, falsehood, and stealing; to be content with one's own wife; and to limit one's possessions. The other sets of vows are supplementary in nature, meant to strengthen and protect the *aṇuvratas*. They involve avoidance of unnecessary travel, harmful activities, and the pursuit of pleasure; fasting and control of diet; offering of gifts and service to monks, the poor, and fellow believers; and voluntary death if the observance of vows proves impossible.

The *sāmāyika*, a lay meditative and renunciatory ritual of limited duration, aims at strengthening equanimity of mind and resolve to pursue the spiritual discipline of the Jaina *dharma*. This ritual brings the lay votary close to the demands required of an ascetic for a limited time. It may be performed in a person's own house, in a temple, in a fasting hall, or before a monk.

Eleven *pratimās*, or stages of a householder's spiritual progress, are listed. Medieval writers conceived *pratimā* (literally, "statue") as a regular progressing series, a ladder leading to higher stages of spiritual development. The last two stages lead logically to renunciation of the world and assumption of the ascetic life.

The disciplines to which Jaina laity must adhere have influenced significantly the types of vocations that they pursue. Since all of their actions should minimize acts of violence to other living creatures, Jainas tend to pursue commercial and professional enterprises and to avoid such careers as military service. This has created an ironic situation in which many adherents to a highly austere and ascetic religion are wealthy.

Sacred times and places. *Festivals and fairs.* The principal Jaina festivals are connected with the five major events in the life of each Tirthankara. These mark the occasions of the Tirthankara's descent into his mother's womb, birth, renunciation, attainment of omniscience, and final emancipation.

The festival of Paryuṣaṇa

The most popular Jaina festival is Paryuṣaṇa, or Paijusaṇa, which occurs in the month of Bhādrapada (August–September). Paryuṣaṇa literally means (1) pacification by forgiving and service with wholehearted effort and devotion and (2) staying at one place for the monsoon season. On the last day of the festival, Jainas distribute alms to the poor and take a Jina image in procession through the streets. Confession is performed during the festival to remove all ill feelings about conscious or unconscious misdeeds during the past year.

Twice a year, for nine days (March–April and September–October), a fasting ceremony known as *olī* is observed. These are also the eight-day festivals corresponding to the mythical celestial worship of images of the Jinas.

On the full-moon day of the month of Kārttika (October–November), at the same time that Hindus celebrate Dewali (festival of lights), Jainas commemorate the Nirvana of Mahāvīra by lighting lamps. Five days later is Jñānapañcamī (literally, "The Fifth Level of Knowledge," *i.e., kevala*), which the Jainas celebrate with temple worship and with worship of the scriptures. Mahāvīra Jayanti, the birth date of Mahāvīra, is celebrated in early April.

The Jainas also celebrate a number of festivals in common with Hindus, such as Holi (spring festival), Navaratra (nine nights festival), and Poṅgal (a South Indian spring festival).

Pilgrimages and shrines. The erection of shrines and the donation of religious manuscripts are regarded as pious acts. Most villages or towns inhabited by Jainas have at least one Jaina shrine; some have become pilgrimage sites. Lists of these shrines have been composed, and the most noteworthy shrines are offered adoration in daily worship.

Places of pilgrimage were created at sites marking the principal events in the lives of Tirthankaras. Parasnāth Hill and Rājgīr in Bihār and Śatruñjaya and Girnār hills on the Kāthiāwār Peninsula are among such important ancient pilgrimage sites. Other shrines that have become pilgrimage destinations are Śravaṇa Belgoḷa in Karnātaka, Mounts Abu and Kesariajī in Rājasthān, and Antarikśa Pārśvanātha in Akola district, Mahārāshtra.

Cave temples

Several Jaina cave temples, dating from as early as the 2nd century BC, have been discovered and excavated. Cave temples are found at Udayagiri and Khandagiri, in Orissa; Rājgīr, in Bihār; Aihole, in Karnātaka; Ellora, in Mahārāshtra; and Sittānnavāsal in Tamil Nādu.

Temple worship and observance. Temple worship is mentioned in early texts that describe gods worshiping Jina images and relics in heavenly eternal shrines. Worship, closely associated with the obligatory rites of the laity, is offered to all liberated souls, to monks, and to the scriptures. Though Tirthankaras remain unaffected by offerings and worship, such actions serve as a form of meditative discipline for the votary offering them. Daily worship includes recitation of the names of the Jinas and idol worship by bathing the image and making offerings to it. Svetambaras decorate images with clothing and ornaments. The worshiper also chants hymns of praise and prayers and mutters sacred formulas. Such Jaina rituals show considerable similarity in form to Hindu rituals. A long-standing debate within both Jaina communities over the centuries has concerned the relative value of external acts of worship and internalized acts of mental discipline and meditation.

Domestic rites and rites of passage. Early Jaina literature is silent about domestic rites and rites of passage marking the main events in a person's life. These rituals are modeled mainly on the 16 Hindu samskaras, which include conception, birth, naming, first meal, tonsure, investiture with the sacred thread, beginning of study, marriage, and death. They are first discussed in Jinasena's 9th-century work, *Ādipurāṇa*.

Welfare institutions. Jainas are renowned for various types of munificence, such as sponsoring pilgrimages, famine relief, relief to Jaina widows and the poor, and maintaining shelters for old animals to save them from slaughter (an act of *ahiṃsā*). In addition, Jainas have encouraged research in and publication of editions of Jaina canonical and commentarial texts. Noteworthy in this connection are the Bhāratīya Jñānapīṭha publishing house in Vāranāsi, Uttar Pradesh, and Lalbhai Dalpatbhai Institute for Indological Research at Ahmadābād, Gujarāt.

JAINA LITERATURE

Canonical and commentarial literature. Jaina canonical scriptures do not belong to a single period, nor is any text free from later revision or additions. The sacred literature, preserved orally from the time of Mahāvīra, was first systematized in a council at Patna about the end of the 4th century BC, and again in two later councils at Mathura and Valabhī in the early 3rd century AD. The fourth and last council, at Valabhī in the mid-4th century, is considered the source of the existing Svetambara canon, though some commentators insist that the present reading is in accordance with the Mathura council.

The original, unadulterated teachings of the Jinas are said to be contained in 14 texts, called the Pūrvas ("Foundation"), which are now lost. Svetambaras and Digambaras agree that a time will come when the teachings of the Jinas will be completely lost; Jainism will then disappear from the earth and reappear at an appropriate point in the next time cycle (*kalpa*). The two sects disagree, however, about the extent to which the corruption and loss of the Jinas' teachings has already occurred. Consequently, the texts for each sect differ.

Svetambara canon

The Svetambaras follow an extensive canon (*āgama*) as the repository of their tradition, which they believe is based upon compilations of Mahāvīra's discourses by his disciples. This canon preserves the teachings of Mahāvīra in an imperfect way, as it is thought to be mixed with much that was not said by the Jina. The number of texts considered to make up the Svetambara canon has varied over time and by monastic group. Largely through the influence of the 19th-century German scholar Johann Georg Bühler, however, Western scholars have fixed the number of texts in this canon at 45, divided into six groups: the 11 Aṅgas ("Parts"; originally there were 12, but one, the Dṛṣṭivāda, has been lost), 12 Upāṅgas (subsidiary texts), four Mūla-sutras (basic texts), six Cheda-sutras (concerned with discipline), two Cūlikā-sutras (appendix texts), and 10 Prakīrṇakas (mixed, assorted texts). The Aṅgas contain several dialogues, mainly between Mahāvīra and his disciple Indrabhūti Gautama, presumably recorded by the disciple Sudharman, who transmitted the teachings to his own disciples.

According to modern scholars, the Ācārāṅga and the Sūtrakṛtaṅga, among the Aṅgas, and the Uttarādhyayana, among the Mūla-sutras, are among the oldest parts of the canon. The Cheda-sutra text, Daśāśrutaskandha, concludes with the Kalpa-sutra, which recounts the lives of the Jinas and includes an appendix of rules for monastic life and a list of eminent monks.

Bhadrabāhu, whom tradition credits with being the last Jaina sage to know the contents of the Pūrvas, is asserted to be the author of the Niryuktis, the earliest commentaries on the Jaina canonical texts. These concise, metrical commentaries, written in Prakrit, gave rise to an expanded corpus comprising texts called Bhāṣyas and Cūrṇis. These were composed between the 4th and 7th centuries and contain many ancient Jaina historical and legendary traditions, along with a large number of popular stories brought into the service of Jaina doctrine. The Bhāṣyas and Cūrṇis, in turn, gave rise in the medieval period to a large collection of Sanskrit commentaries. Haribhadra, Sīlāṅka, Abhayadeva, and Malayagiri are the best-known authors of such commentaries.

Digambara canon

Digambaras give canonical status to two works in Prakrit: the Karmaprābhṛta ("Chapters on Karman," also called Ṣhaṭkhaṇḍāgama) and the Kaṣāyaprābhṛta ("Chapters on the Kaṣāyas"). The Karmaprābhṛta, based on the now-lost Dṛṣṭivāda text, deals with the doctrine of *karman* and was committed to writing by Pushpadanta and Bhūtabalin in the mid-2nd century; the Kaṣāyaprābhṛta, compiled by Guṇadhara from the same source at about the same time, deals with the passions that defile and bind the soul. Later commentaries by Vīrasena (8th century) and his disciple Jinasena (9th century) on the Kaṣāyaprābhṛta are also highly respected by Digambaras.

Philosophical and other literature. In addition to the canons and commentaries, the Svetambara and Digambara traditions have produced a voluminous corpus of literature, written in several languages, in the areas of philosophy, poetry, drama, grammar, music, mathematics, medicine, astronomy, astrology, and architecture. In Tamil, the epics *Cilappatikāram* and *Jīvikacintāmaṇi,* which are written from a Jaina perspective, are important works of early postclassical Tamil literature. Jaina authors were also an important formative influence on Kannada literature. The *Ādipurāṇa* of the Jaina lay poet Pampa (another text dealing with the lives of Ṛṣabha, Bāhubali, and Bharata) is the earliest extant piece of *mahākāvya* ("high poetic") Kannada literature. Jainas were similarly influential in the Prakrit languages, Apabhramsa, Old Gujarati, and, later, Sanskrit. A particularly important literary figure in Prakrit and Sanskrit was the Svetambara monk Hemacandra (12th century), who composed an important Prakrit grammar, as well as poetry, philosophical treatises, and a mammoth epic poem on the lives of the 63 Jaina *mahāpuruṣa*s, entitled *Triṣaṣṭiśalākāpuruṣacaritra.*

Other noncanonical Jaina writers on philosophy include Mallavādin I (4th century), Siddhasena Divākara (*c.* 5th century), Haribhadra Sūri (*c.* 8th century), Samantabhadra (before the 5th century), Akalaṅka (*c.* 8th century), Siddharṣi Gaṇin (10th century), Śāntisūri (11th century), Vidyānandin (*c.* 8th–9th century), Anantakīrti (10th century), Māṇikyanandin (11th century), Prabhācandra (11th century), and Vādi Deva Sūri (12th century). Among later authors, UpādhyāyaYaśovijaya (*c.* 17th century), a versatile scholar, is especially noteworthy.

Digambaras also value the Prakrit works of Kuṇḍakuṇḍa (*c.* 2nd century), including the *Pravacanasāra* (on ethics), the *Samayasāra* (on fine entities), the *Niyamasāra* (on Jaina monastic discipline), and the six *Prābhṛta*s ("Chapters") on various religious topics. Of similar importance is the *Tattvārthādhigama-sutra* of Umāsvāmin (or Umāsvāti), whose work is claimed by both communities. Composed early in the Christian Era, the *Tattvārthādhigama-sutra* was the first work in Sanskrit on Jaina philosophy dealing with such subjects as logic, epistemology, ontology, ethics, cosmography, and cosmogony; it generated numerous commentaries, including one by Umāsvāti himself.

RELIGIOUS SYMBOLISM AND ICONOGRAPHY

Image worship was introduced at an early stage, perhaps even during the century immediately following the death of Mahāvīra. The Jina himself appears to have made no statement regarding the worship of images. Descriptions of *stūpas* (reliquaries for the bones and ashes of saints), commemorative pillars, and tree shrines appear in early Jaina texts, which also refer to the worship in the heavens by gods of images of the four legendary Śāśvata Jinas ("Eternal Victors") and of costly relic boxes. Mention is made of *śilāpaṭa*s, which apparently were stone plaques or reliefs placed on lion thrones underneath trees, such as those associated with the worship of Yakshas (mythical nature spirits), and also depicted on Buddhist reliefs from Bharhut (2nd century BC). The *śilāpaṭa*s appear to be the prototypes of the later Jaina *āyāgapaṭa*s (tablets of homage) from Mathura (Uttar Pradesh state), which show representations of *stūpa*s, *caitya* pillars surmounted by elephants, *dharmacakra*s (wheels of the law), and the *aṣṭamaṅgala*s (eight auspicious symbols). Later *āyāgapaṭa*s show a Jina attended by two nude disciples and the figure of the monk Kaṇha Samaṇa with his disciples, or they depict the figure of a noblewoman with attendants.

Āyāgapaṭas

The earliest extant Tirthankara image is possibly the highly-polished Mauryan period torso from Lohanipur, near Patna. Numerous Tirthankara images in the sitting and standing postures dating from the early Christian Era have been uncovered in excavations of a Jaina *stūpa* at Mathura. The earliest images of Tirthankaras are all nude. The various Jinas are distinguished by inscriptions giving their names carved on the pedestals, but later iconographic devices such as symbols specific to each Jina did not evolve until about the 5th century.

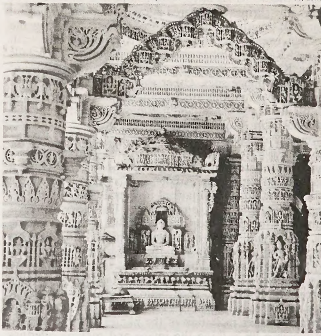
Worship of the 16 principal Jaina Tantric goddesses, the Mahāvidyās, was probably introduced in the Gupta age. From the 6th to the 11th century a common pair of attendants was employed in sculpture for all the Tirthankaras, but from about the 9th century 24 *śāsanadevatā*s were evolved, each one to attend a different Tirthankara. The names of many of the attendants suggest Hindu or Buddhist influence.

The religious merit that accrues from hearing and reading Jaina texts encouraged the careful and loving preservation of illustrated manuscripts. The miniature paintings on palm-leaf and paper manuscripts preserved in the Jaina monastic libraries provide a continuous history of the art of painting in western India from the 11th century to the present. The lives of the Jinas and legends of Jaina saints provide a framework for the artists to depict gods and goddesses, throne rooms and village interiors, gardens, and temples. Religious symbols such as the *ashṭamaṅgala*s and the 14 dreams of the mothers of the Tirthankaras frequently appear in paintings.

In addition to the miniatures and to painted wooden book covers that often show mythological scenes, paintings on cloth are also known. Wall paintings are found on cave shrines at Sittānavāsal (Tamil Nādu state) and at Ellora.

Jaina temples generally contain a number of metal images of various types and metal plaques showing auspicious symbols. Metal images of the Jinas are also kept by pious Jainas for home devotion. Among the earliest known bronzes are one of Pārśvanātha in the Prince of Wales Museum of Western India in Bombay, which may date from the 1st century BC, and a group of bronzes (1st–3rd century AD) from Chausa in Bihār in the Patna Museum.

By courtesy of the Government of India Tourist Office, London



Interior of a Jaina temple, showing the Tirthankara image enshrined. Dilwāra Temple, Mount Abu, India, 13th century AD.

JAINISM AND OTHER RELIGIONS

Jainism, Hinduism, and Buddhism. Jainism, Hinduism, and Buddhism share a discourse made available through the Sanskrit language and the dialects (Prakrits) derived from it. Having a set of key concepts in common has enabled these traditions to finely hone their religious debates. For example, all three traditions share a notion of *karman* as the actions of individuals that determine their future births; yet, each has attached connotations to the concept that are uniquely its own. This is also true with terms such as *dharma* (often translated "duty," "righteousness," or "religious path"), *yoga* ("ascetic discipline"), and *yajña* ("sacrifice," or "worship"). This Sanskritic discourse has been brought into the service of the religious and philosophical speculations, as well as the polemics, of each of these traditions.

Shared key concepts

The same circumstance occurs in the ritual life and literature of each religion. In the ritual sphere, for example, the *abhiṣeka,* or head-anointing ritual, has had great significance among all three, especially in royal contexts. The best-known example of this ritual is the one performed every 12 to 14 years on the statue of Bāhubali at the Jaina pilgrimage site at Śravaṇa Beḷgoḷa. The structure of this ritual is similar in each religious context; in each case, however, it has specific meanings peculiar to that context.

In the literary sphere, each tradition developed an extensive corpus of canonical and commentarial literature, and each has developed a body of narrative literature. For example, so great was the influence of the story of Rāma in the classical Hindu *Rāmāyaṇa,* that the Buddhists and Jainas felt obliged to retell the story in their own terms. Jaina literature includes 16 different tellings of this story in Sanskrit and Prakrit.

Finally, each tradition shares a similar understanding of the ascetic life, though each understands it as functioning properly only within the context of its own religious system. Many of the terms applied to figures in each monastic organization are the same (though not necessarily the same in meaning), and several of the monastic ritual and meditative activities are similar in structure.

Jainism and Islām. In reference to Muslim influence on Jainism, it has been suggested that the concept of *āśātanās*—activities that are unfitting or indecent in a temple—reveals a notion of the sanctity of the temple that is more evocative of Muslim *barakah* ("holiness") than of any traditional Jaina attitude. The most obvious influence of Islām is seen, however, in the repudiation by the Svetambara Loṅkāsāha sect of image worship as something without canonical support. A parallel sect, the Terāpanthin, also arose among the Digambaras.

Jaina influence at the Mughal court of Akbar is a bright chapter in Jaina history. Akbar honoured Hīravijaya Sūri, then the leader of the Svetambara Tapā *gaccha* (subgroup). His disciples and other monks gained the respect of the Mughal emperors Jahāngīr, Shāh Jahān, and even the Muslim chauvinist Aurangzeb. Akbar issued a decree prohibiting animal slaughter near important Jaina sites during the Paryuṣaṇa festival. Jahāngīr also issued decrees for the protection of Śatruñjaya, and Aurangzeb issued a decree favouring the Jainas with respect to proprietary rights over Mount Śatruñjaya. Mughal painting, influential in different schools of Indian painting, also influenced Jaina miniature painting.

BIBLIOGRAPHY

General sources: Good introductions are HERMANN JACOBI, "Jainism," in *Encyclopædia of Religion and Ethics,* vol. 7, pp. 465–474 (1928); and COLETTE CAILLAT, "Jainism," in *The Encyclopedia of Religion,* ed. by MIRCEA ELIADE, vol. 7, pp. 507–514 (1987). Standard works include HERMANN JACOBI (trans.), *Gaina Sūtras,* 2 vol. (1884–95, reissued as *Jaina Sutras,* 1968), with noteworthy introductions by Jacobi to each volume; JOHANN GEORGE BUHLER, *On the Indian Sect of the Jainas,* trans. from German, 2nd ed. (1963); HELMUTH VON GLASENAPP, *Der Jainismus* (1925, reprinted 1964), the most comprehensive text on Jainism, and *The Doctrine of Karman in Jain Philosophy,* trans. from German (1942); and WALTHER SCHUBRING, *The Doctrine of the Jainas* (1962; originally published in German, 1935), a scholarly work, and *The Religion of the Jainas,* trans. from German (1966). See also CHHOTELAL JAIN, *Chhotelal Jain's Jaina Bibliography,* 2nd. rev. ed., edited by SATYA RANJAN BANNERJEE, 2 vol. (1982); AMULYACHANDRA SEN, *Schools and Sects in Jaina Literature* (1931); JAGMANDERLAL JAINI, *Outlines of Jainism* (1916, reprinted 1982); A.L. BASHAM, *History and Doctrine of the Ājīvikas* (1951, reprinted 1981), a discussion of the Ājīvika influence on early Jainism; CHHOGMAL CHOPRHA, *A Short History of the Terapanthi Sect of the Swetamber Jains and Its Tenets,* 4th ed. (1950); BIMALA CHURN LAW, *Mahavira: His Life and Teachings* (1937), a good introduction to the subject; and PADMANABH S. JAINI, *The Jaina Path of Purification* (1979), a survey that discusses the Jaina understanding of karmic bondage and the path to liberation.

Special studies: NATHMAL TATIA, *Studies in Jaina Philosophy* (1951, reprinted 1980), especially the discussion on the problem of *ajñāna,* or false sense of reality, in various systems; NARENDRA NATH BHATTACHARYYA, *Jain Philosophy: Historical Outline* (1976); SATKARI MOOKERJEE, *The Jaina Philosophy of Non-Absolutism: A Critical Study of Anekāntavāda,* 2nd ed. (1978), a standard work by an authority on Indian philosophy; MOHANLAL MEHTA, *Jaina Philosophy,* new ed. (1971), *Jaina Culture* (1969), and *Jaina Psychology: A Psychological Analysis of the Jaina Doctrine of Karma* (1957); SHANTARAM B. DEO, *History of Jaina Monachism from Inscriptions and Literature* (1956); R. WILLIAMS, *Jaina Yoga* (1963, reprinted 1983), a masterly analysis of the Jaina ethics concerning the laity, with critical notes on authors of different sourcebooks; DAYANAND BHARGAVA, *Jaina Ethics* (1968); HARI SATYA BHATTACHARYA, *Jain Moral Doctrine* (1976); T.K. TUKOL, *Sallekhanā Is Not Suicide* (1976), a treatise on the monastic ritual of self-starvation; KAMAL C. SOGANI, *Ethical Doctrines in Jainism* (1967); VILAS ADINATH SANGAVE, *Jaina Community: A Social Survey,* 2nd ed. (1980); CHAMPAT R. JAIN, *Jaina Law* (1926); COLETTE CAILLAT, *Attonements in the Ancient Ritual of the Jaina Monks* (1975; originally published in French, 1965); COLETTE CAILLAT and RAVI KUMAR, *The Jain Cosmology* (1981); and A.N. UPADHYE, *Upadhye Papers* (1983), a collection of essays on Jaina history and literature by an eminent Jaina scholar.

Literature and art: M. WINTERNITZ, *History of Indian Literature,* vol. 2 (1933, reprinted 1971; originally published in German, 1920); H.R. KAPADIA, *A History of the Canonical Literature of the Jainas* (1941), a good description of the Jaina canon; A. CHAKRAVARTY, *Jaina Literature in Tamil* (1974), a survey of Jaina works in this South Indian language and Jaina influence on Tamil literature; JAGDISHCHANDRA JAIN, *Prakrit Narrative Literature: Origin and Growth* (1981); B.C. BHATTACHARYA, *The Jaina Iconography,* 2nd rev. ed. (1974), a brief outline of the subject; JYOTINDRA JAIN and EBERHARD FISCHER, *Jaina Iconography,* 2 vol. (1978), a later work; UMAKANT P. SHAH, *Studies in Jaina Art* (1955), a review of Jaina art in North India, with a discussion of various symbols in Jaina worship and a good bibliography, and *Akota Bronzes* (1959), a description of rare Jaina bronzes from a site in Gujarāt; A. GHOSH (ed.), *Jaina Art and Architecture,* 3 vol. (1974–75); KLAUS FISCHER, *Caves and Temples of the Jains* (1956); MOTI CHANDRA, *Jain Miniature Paintings from Western India* (1949), a standard textbook; W. NORMAN BROWN, *The Story of Kālaka* (1933), a well-known work on Kālakācārya and miniature Jaina paintings; UMAKANT P. SHAH (ed.), *Treasures of Jaina Bhaṇḍāras* (1978); and P.B. DESAI, *Jainism in South India and Some Jaina Epigraphs* (1957), a useful compilation.

(U.P.S./G.R.S.)

Jakarta

Jakarta is the capital of the Republic of Indonesia and one of the largest and most consistently growing cities in that country. From the early 17th century, when Dutch colonizers destroyed and subsequently rebuilt the settlement, until 1949, the city was called Batavia; its present name was adopted in that year but was spelled Djakarta until 1972. Coextensive with the metropolitan district of Jakarta Raya, it has an area of 256 square miles (664 square kilometres) and lies at the mouth of the Ciliwung (Liwung River) on the northwest coast of Java.

In 1966 the city was declared to be a special metropolitan district (*daerah khusus ibukota*), thus gaining a status approximately equivalent to that of a state or province. The city has long been a major trade and financial centre; it has also become an important industrial city and an important centre for education.

This article is divided into the following sections:

Physical and human geography

THE LANDSCAPE

The city site. Jakarta lies on a low, flat alluvial plain, with extensive swampy areas; it is easily flooded during the rainy seasons. The parts of the city farther inland are slightly higher. The draining of swamps for building purposes and the continuous decrease of upland forest vegetation have increased the danger of floods. With an excess of water in the soil, Jakarta still has a shortage of clean drinking water, for which there is an increasing demand. The area is quite fertile for fruit and other horticulture, as most of the soil is of old volcanic origin.

Climate. Jakarta is a tropical, humid city, with annual temperatures ranging between the extremes of 75° and 93° F (24° and 34° C) and a relative humidity between 75 and 85 percent. The average mean temperatures are 79° F (26° C) in January and 82° F (28° C) in October. The annual rainfall is more than 67 inches (1,700 millimetres). Temperatures are often modified by sea winds. Jakarta, like any other large city, now also has its share of air and noise pollution.

The city layout. Although the Dutch were the first to attempt to plan the city, the city layout is probably more British than Dutch in character, as can be seen from such large squares as the Medan Merdeka (Freedom Field) and Lapangan Banteng (Place of the Gaur [large wild ox]). The Oriental style, or "indische" style, as the Dutch call it, is, however, apparent not only in the city's way of life but also in the types of houses, the wide, tree-lined streets, and the original spacious gardens and house lots. In Kebayoran, a satellite town built since World War II on the southwestern side of the city, and in other modern developments, the houses and garden lots are much smaller than in the older colonial districts.

Traditional neighbourhoods

Jakarta has long been a city of new settlers who assimilated local ways and became Jakartans themselves. Some traditional neighbourhoods can, however, be identified. The Kota (Fort), or Old City, for example, sometimes called the downtown section, is the central business district and Indonesia's financial capital. It houses a significant part of the Chinese population. The area of Kemayoran (Progress) and Senen, originally on the eastern fringe of the city, is now almost central in its location and increasingly has become the city's major retail area.

The Jatinegara (Real Country) section, originally a Sundanese settlement but later incorporated as a separate town, then a Dutch army camp (Meester Cornelis), has now merged with the rest of Jakarta and includes many new settlers. The Menteng and Gondangdia sections were formerly fashionable residential areas near the central Medan Merdeka (then called Weltevreden). To the west, Tanahabang (Red Earth) and Jati Petamburan are, like Kemayoran, densely developed. Tanjungpriok is the harbour, with its own community attached to it.

The most common type of house in the city is the kampong, or village house. Most are built of materials such as wood or bamboo mats, but this does not necessarily mean that this is substandard. Another common type of housing, often used to house government workers, is the colonial

H. Armstrong Roberts

Presidential Palace, Jakarta.

Jakarta and (inset) its metropolitan area.

urban house, or *rumah gedongan;* these are mostly single-family detached or semidetached houses, each standing on a separate lot. Apartment buildings constitute a more modern category; although they are more economical in the use of land than single-family types, their architectural and construction costs often make them fairly expensive. Housing is generally overcrowded.

Some of Jakarta's buildings, such as the Portuguese Church (1695) in the Old City, are of architectural or historical interest. Some of the buildings around the city square in the Kota also date from colonial times, including the old city hall (1710), which has been restored and now serves as the municipal museum. The National Archives building was originally the palace of a Dutch governor general, Abraham van Riebeeck. The Ministry of Finance building, facing Banteng Square, also was designed as a governor's palace (Herman Willem Daendels, one of Napoleon's marshals). The Presidential Palace, north of

Buildings from colonial times

Medan Merdeka, faces the Monas, or Monumen Nasional (National Monument), at 360 feet (110 metres) the highest building in Jakarta. The Istiqlal Mosque, in the northeast corner of Medan Merdeka opposite Banteng Square, is one of the largest mosques in Southeast Asia. The National Museum (formerly the Central Museum), on the west side of the Medan Merdeka, houses a collection of historical, cultural, and artistic artifacts.

After World War II Jakarta underwent a building boom. The Hotel Indonesia (the city's first high-rise building) and the Senayan Sports Complex were built for the Asian Games in 1962. Most high-rise buildings are located along Husni Thamrin and Jendral Sudirman roads, connecting Jakarta with Kebayoran.

THE PEOPLE

The population of Jakarta has increased dramatically since 1940. Much of this increase is attributed to immigration, which has transformed Jakarta into one of the world's largest urban agglomerations. Although government regulations close the city to unemployed new settlers, better economic conditions inevitably attract new people. In addition, much of the population is young, resulting in a very high natural increase potential. Analysis of the immigrant stream shows that after the West Javanese, the largest groups represented are the Central and East Javanese; a sizable number also are from Sumatra. Other population groups—Arabs, Indians, Europeans, and Americans—are present in small numbers.

THE ECONOMY

Economically, Jakarta plays several roles. It can be identified first as the national capital and a central place of control for the national economy, then as an administrative centre in its own right, and, finally, as a significant industrial hub. In addition, its location as a port makes it an important centre for trade.

Industry. Jakarta has some manufacturing industries, including several iron foundries and repair shops, margarine and soap factories, breweries, and printing works. Machinery, cigarettes, paper, glassware, wire cable, and aluminum and asbestos, and more recently also automotive, products are manufactured. There are also tanneries, sawmills, textile mills, food-processing plants, and a film industry.

Commerce and trade. The cost of living in the city continues to rise. Land is expensive, and rents are high, so that industrial development and the construction of new housing usually are undertaken on the outskirts, while commerce and banking remain concentrated in the city centre. The Indonesian Chamber of Commerce is active in promoting trade with other countries; the annual Jakarta Fair (usually held from July to August) also serves to promote trade. Jakarta is the centre of roughly one-fourth of Indonesia's trade and services and two-thirds of its banking and financial sectors.

City-operated markets

To meet the needs of the local city population, the municipality operates several markets. The central city markets (Pasar Kota), like the markets of Pasar Senen to the east of the central city and Pasar Glodok in the Kota area, are major retail centres. The Pasar Jatinegara is primarily a food supply centre. The district markets are fairly large, with each one catering to a whole section of the city. There are also small neighbourhood markets, each serving only a limited area. Special markets include one selling fish, one selling used and new automobile parts, the Pasar Rumput flea market, and the Jalan Surabaya souvenir and antique market. Jakarta also has several general neighbourhood markets.

Transportation. Major road arteries lead west from the centre of the Kota Old City and east and south from the administrative centre in Gambir. Traffic congestion is a major problem in the city, as is pollution. To the east, a major railroad connects the city with all of the island of Java. There is also a highway, primarily a regional supply road, running between Jakarta and the agriculturally productive areas of East and Central Java. To the south, a road and railroad connect Jakarta with Bogor, Sukabumi, and Bandung. To the west, a railroad and road run to Banten and to the harbour in Merak, which is connected by ferry to Lampung in Sumatra.

The port of Tanjungpriok in Jakarta is the largest in Indonesia, handling exports from West Java and a large proportion of Indonesia's import trade; many goods are transshipped to other islands or harbours.

Jakarta is served by several international airlines, by Garuda Indonesian Airways (the national airline, with international and domestic service), and by other domestic airlines. The city's main international airport is located about 12 miles (20 kilometres) to the west at Cengkareng, and a smaller facility is just to the southeast.

The central bus terminal, located on Lapangan Banteng, serves all the city, intercity, and regional bus lines; there are also suburban bus terminals in Jatinegara, Kebayoran, Grogol, Kota, and Tanjungpriok. The major railroad stations are at Kota in the old city, Gambir on Medan Merdeka, Pasar Senen on the east, and Manggarai and Jatinegara on the south. Tanahabang serves the west and traffic to Merak. Traffic jams occur particularly during the morning and afternoon rush hours. Public transportation in the city is by bus or minibus. The *becak,* or tricycle taxi, is used only for local neighbourhood transportation, and regular taxis now operate throughout the metropolitan area.

The *becak,* or tricycle taxi

ADMINISTRATION AND SOCIAL CONDITIONS

Government. Jakarta is formally classified as a metropolitan district. The city is thus at the administrative level of an Indonesian province, and its mayor has the same status as the governor of a province. Jakarta itself is divided into five administrative municipalities. The mayor is responsible for supervising the administration of local government and for coordinating the local offices of national government ministries. The city government is composed of two branches, the executive and the electorate. The executive consists of a governor assisted by four vice governors, an executive staff, and a regional secretary; there are also a number of city directorates, bureaus, and agencies attached to the executive. The electorate consists of select members drawn from political parties, the armed forces, and so-called functional groups. It is headed by a council of five members, one chairman, and four vice chairmen.

Public utilities. Public utilities are usually operated or owned by the Indonesian government. The State Electricity Company and the subsidiary State Gas Company both supply Jakarta. Postal and cable services and telephone services are supplied by state companies working under the aegis of the Department of Communications. Jakarta's electricity comes from several sources; these include the thermal plant in Ancol, close to the port of Tanjungpriok, smaller diesel plants in various parts of the city, and the Jatiluhur hydroelectricity project located close to Purwakarta, about 70 miles southeast of Jakarta. A thermal power plant operates in Surabaya.

The city government is responsible for the water supply. The city water is obtained in part from freshwater springs in the Bogor area, but most of the supply comes from the Pejompongan water treatment plant. Water is required mainly for domestic purposes but is also needed for industry and to supply ships. The removal of garbage and the provision of other sanitation services are also the responsibility of the municipality.

Health. Jakarta is the centre for many health facilities in Indonesia. There are several major hospitals, including those operated directly by the Department of Health, the Roman Catholic church, and by a Protestant mission. Municipal hospitals serve separate areas of the city. Altogether Jakarta has about 40 general or special hospitals. In addition, there are several hundred general clinics or polyclinics located throughout the city. A quarantine hospital is in operation in Tanjungpriok. The city also operates a hospital and rehabilitation centre for the mentally ill and destitute, and there are many family-planning and child care clinics.

Hospitals and clinics

Education. To meet the needs for primary education, many new elementary schools and secondary schools have been built, and a number of old school buildings have been renovated. There is a well-developed system of kindergartens, elementary schools, *madrasahs* (religious schools), secondary schools, and high schools. There are also many vocational and special schools and more than 100 universities, academies, and institutes for higher learning. The largest and best known university is the Universitas Indonesia (founded 1950).

CULTURAL LIFE

Among other cultural activities, the Taman Ismail Marzuki centre has facilities for traditional or classical art performances as well as theatres for presenting modern plays and concerts; the centre also has a planetarium. Traditional performances include *wayang* dance and drama, gamelan music, and *wayang* puppet shows. Traditional performances representing the culture of other parts of Indonesia are included in the programs presented at the annual Jakarta Fair.

Extensive public recreation areas in and around Jakarta include a seaside recreation area at Ancol and the Ragunan Zoological Gardens, near Pasarminggu. The latter houses such indigenous animals as Komodo dragons, Sumatran tigers, and orangutans. Playgrounds include, among others, the Taman Ria and the Taman Ria Remaja complexes. The fair grounds, designed as an exhibition facility, opened in 1992. Jakarta's historic section, known as "Old Batavia," highlights the city's Dutch influence. Although some of the buildings there were destroyed in the 19th century, a major restoration project got underway in the 1970s. The 250-acre (100-hectare) Taman Mini Indonesia Indah (Beautiful Indonesia in Miniature) park, just southeast of the city, contains exhibits of traditional houses representing each of Indonesia's provinces. The city also provides public recreation facilities.

History

The first settlements on the site of Jakarta were established at the mouth of the Ciliwung, perhaps as early as the 5th century AD. The city's official history, however, starts in 1527, when the sultan of Bantam defeated the Portuguese there and called the place Jayakerta (Sundanese: "Glorious Fortress").

The Dutch, under the leadership of Jan Pieterszoon Coen, captured and razed the city in 1619, after which the capital of the Dutch East Indies—a walled township named Batavia—was established on the site.

The colonial era

The colonial history of the city can be divided into three major periods. First was that of the Dutch East India Company, when most of the activities of the city centred around the fortress and the company warehouses. At that time the city somewhat resembled a typical Dutch town, complete with canals. The second period began in the early 1800s, when the city was extended to include higher and more healthful areas to the south, which would later become the seat of the new colonial government. A brief interval of British control during the Napoleonic Wars, ending in 1815, interrupted the second period. During the third period, which lasted from about the 1920s to 1941, the city became modernized.

The colonial era ended with the entry of Japan into World War II, when Indonesia was occupied by Japanese forces. After the war the city was briefly occupied by the Allies and then was returned to the Dutch. During the Japanese occupation and again after Indonesian nationalists declared independence on August 17, 1945, the city was renamed Djakarta. The Dutch name Batavia remained the internationally recognized name until full Indonesian independence was achieved and Djakarta was officially proclaimed the national capital (and its present name recognized) on December 27, 1949.

Jakarta has undergone tremendous growth and development since Indonesia's independence. During the 1950s the city began its transformation, as President Sukarno supported large-scale construction projects. The city's infrastructure was modernized, and office towers rose. Jakarta became one of the largest metropolises of tropical Asia and emerged as a financial and commercial centre during the tenure of Suharto, whose reign was marked by widespread corruption and nepotism. After decades of economic growth, the city and the country were hit hard by a financial crisis in Asia in 1997–98. Widespread antigovernment protests and rioting ensued, which claimed the lives of more than 500 people; Suharto was forced to resign in 1998. Thus, at the outset of the 21st century Indonesia and Jakarta were undergoing a dramatic economic, political, and social transition.

BIBLIOGRAPHY. On the history of Jakarta, see F. DE HAAN, *Uit Oud-Batavia* (1898); DJAKARTA, KOTAPRADJA, *Sedjarah pemerintahan kota Djakarta* (1958); and SUSAN ABEYASEKERE, *Jakarta: A History,* rev. ed. (1989). Census and population information may be found in the Jakarta volumes issued by the BIRO PUSAT STATISTIK in its *Sensus penduduk* (Census of Population) reports for the censuses of 1961, 1971, 1980, and 1990; and in H.J. HEEREN (ed.), *The Urbanisation of Djakarta* (1955). See also *Djakarta: Its Rehabilitation and Development* (n.d.), issued by the Badan Perentjana Pembangunan (Development Planning Body) of the City Government; and PAULINE D. MILONE, *Urban Areas in Indonesia* (1966). (W.J.W./Ed.)

Japan

The country of Japan (Japanese: Nihon or Nippon) lies off the east coast of Asia. It consists of a great string of islands in a northeast-southwest arc that stretches for approximately 1,500 miles (2,400 kilometres) through the western Pacific Ocean. Japan has a total land area of 145,898 square miles (377,873 square kilometres). Nearly this entire area is taken up by the country's four main islands; from north to south these are Hokkaido (Hokkaidō), Honshu (Honshū), Shikoku, and Kyushu (Kyūshū). Honshu is the largest of the four, followed in size by Hokkaido, Kyushu, and Shikoku. In addition, there are numerous smaller islands, the major groups of which are the Ryukyu (Nansei) Islands (including the island of Okinawa) to the south and west of Kyushu and the Izu, Bonin (Ogsawara), and Volcano (Kazan) islands to the south and east of central Honshu. The national capital, Tokyo (Tōkyō), in east-central Honshu, is one of the world's most populous cities.

Japan is bounded to the west by the Sea of Japan (East Sea), which separates it from the eastern shores of South and North Korea and southeastern Siberia (Russia); to the north by La Perouse (Sōya) Strait, separating it from Russian-held Sakhalin Island, and by the Sea of Okhotsk; to the northeast by the southern Kuril Islands (under Russian administration); to the east and south by the Pacific; and to the southwest by the East China Sea, which separates it from China. The island of Tsushima lies between northwestern Kyushu and southeastern South Korea and defines the Korea Strait on the Korean side and the Tsushima Strait on the Japanese side.

Complexity and contrast are the keynotes of life in Japan—a nation possessing an intricate and ancient cultural tradition yet one that, since World War II, has emerged as one of the world's most economically and technologically advanced societies. (Ak.W./G.La.)

This article is divided into the following sections:

PHYSICAL AND HUMAN GEOGRAPHY

The land

RELIEF

Mountainous character of Japan

The Japanese landscape is rugged, with more than four-fifths of the land surface consisting of mountains. The mountainous character of the country is the outcome of orogenic (mountain-building) forces largely during the Quaternary Period (the past 1.6 million years), as evidenced by the frequent occurrence of violent earthquakes, volcanic activity, and signs of change in sea levels along the coast. There are no sizable structural plains and peneplains (large land areas leveled by erosion), features that usually occur in more stable regions of the Earth. The mountains are for the most part in a youthful stage of dissection in which steep slopes are incised by dense river-valley networks. Rivers are mostly torrential, and their valleys are accompanied by series of river terraces that are the result of movements in the Earth's crust, as well as climatic and sea-level changes in Holocene times (*i.e.,* the past 10,000 years). Recent volcanoes are juxtaposed with old and highly dissected ones. The shores are characterized by elevated and depressed features such as headlands and bays, which display an incipient stage of development.

The mountains are divided into many small land blocks that are separated by lowlands or deep saddles; there is no long or continuous mountain range. These land blocks are the result of intense faulting and warping and are often bounded by fault scarps and flexure slopes that descend in step formation to the adjacent lowlands.

Coalescing alluvial fans—cone-shaped deposits of alluvium that run together—are formed where rivers emerge from the mountains. When the rivers are large enough to extend their courses to the sea, low deltaic plains develop in front of the fans; this occurs most frequently where the rivers empty into shallow and sheltered bays, as in the deltas of Kantō (Kwanto), Nōbi, and Ōsaka. In most places, however, fan surfaces plunge directly into the sea and are separated by low, sandy beach ridges.

Dissected plains are common. Intense disturbances have caused many former alluvial fans, deltas, and sea bottoms to be substantially uplifted to form flat-topped uplands such as those found in the Kantō Plain. Frequently the uplands have been overlain with volcanic ash, as in the Kantō and Tokachi plains.

Geologic framework. Japan is one of the world's most geologically unstable areas. The country experiences some 1,000 tremors annually, most of them minor, though major quakes—such as in Tokyo-Yokohama (1923) and Kōbe (1995)—cause considerable loss of life and widespread destruction. Violent volcanic eruptions occur frequently, and at least 60 volcanoes have been active within historical time. Among the major eruptions since 1980 are those of Mounts O (1983) and Mihara (1986) in the Izu Islands and Mount Unzen (1991) in Kyushu. The country's abundant hot springs are mostly of volcanic origin. Many of the gigantic volcanoes are conical in shape (*e.g.,* Mount Fuji), while others form steep lava domes (*e.g.,* Mounts Dai and Unzen). Conspicuous shield volcanoes (broad, gently sloping volcanic cones) are rare, and extensive lava plateaus are lacking. One of the characteristics of the volcanic areas is the prevalence of calderas, especially in the northeast and southwest, many of which are filled with water, such as Lakes Kutcharo, Towada, and Ashi.

Tectonic activity

The cause of this instability is the tectonic movement of several of the Earth's major crustal plates in the vicinity of the archipelago. Most important is the subduction (sinking) of the Pacific Plate (in the north) and the Philippine Plate (in the south) beneath the Eurasian Plate upon which Japan lies. The movements of these plates have formed six mountain arcs off the northeastern coast of Asia: from northeast to southwest, the Chishima Range of the Kuril Islands; the Karafuto (Sakhalin) Mountain system of Hokkaido; the Northeast, Southwest, and Shichito-Mariana ranges of Japan; and the Ryukyu formations.

The major physiographic regions. These mountain arcs, in turn, generally correspond to Japan's major physiographic regions: the four regions of Japan proper (Hondo)—Hokkaido, Northeastern, Central, and Southwestern—and the Ryukyu and Bonin archipelagoes.

The Hokkaido region was formed by the coalescence of the north–south-trending Chishima and Karafuto arcs. The Chishima arc enters Hokkaido as three volcanic chains with elevations of more than 6,000 feet (1,800 metres); these are arranged in ladder formation and terminate in the heart of the region. Chief components of the mountain system are the Kitami Mountains in the north and the Hidaka Range in the south.

The Northeastern Region stretches from southwest Hokkaido to central Honshu. Several rows of mountains, lowlands, and volcanic zones are closely oriented to the general trend of the insular arc of this region, which is convex toward the Pacific Ocean. The Kitakami and Abukuma ranges on the east coast are somewhat oblique to the general trend; they are chiefly composed of older rocks, and plateaulike landforms survive in the centre. In the western zone, the formations conform to the general trend and are composed of a basement complex overlain by thick accumulations of young rocks that have been subjected to mild folding. The Ōu Range, capped with towering volcanoes that form the main part of the East Japan Volcanic Belt, is separated from the coastal ranges by the Kitakami-Abukuma lowlands to the east and by a row of basins in the west.

The Central Region of Honshu is dominated by the coalescence of the Northeast, Southwest, and Shichito-Mariana mountain arcs near Mount Fuji, which, at an elevation of 12,388 feet (3,776 metres), is Japan's highest mountain. The trend of the mountains, lowlands, and volcanic zones intersects the island almost at right angles. The most notable physical feature is the Fossa Magna, a great rift lowland that traverses the widest portion of Honshu from the Sea of Japan to the Pacific. It is partially occupied by mountains and volcanoes of the southern part of the East Japan Volcanic Belt. Intermontane basins are sandwiched between the lofty, partially glaciated central mountain knots of the Akaishi, Kiso, and Hida ranges (which together form the Japanese Alps) to the west and the Kantō Range to the east. The shallow structural basin of the Kantō Plain, which stretches to the east of the Kantō Range, is the most extensive lowland of Japan; the immense metropolis of Tokyo spreads out from its centre.

The Kantō Plain

The Southwestern Region of southern Honshu, Shikoku, and northern Kyushu generally coincides with the southwestern mountain arc, and the general trend of highlands and lowlands is roughly convex toward the Sea of Japan. The region is divided into the Inner Zone, formed by complex faulting, and the Outer Zone, formed by warping. The Inner Zone is chiefly composed of granite, Paleozoic (245 to 540 million years old), and volcanic rocks, which are arranged in complicated juxtaposition. The Outer Zone, consisting of the Akaishi, Kii, Shikoku, and Kyushu mountain groups, in contrast, is characterized by a regular zonal arrangement from north to south of crystalline schists and Paleozoic, Mesozoic (66.4 to 245 million years old), and Tertiary (1.6 to 66.4 million years old) formations. The outstanding surface features of the Inner Zone present a highly complex mosaic of numerous fault blocks, while those of the Outer Zone are continuous except where the sea straits separate them into the four independent groups. The Inland Sea (Seto-naikai) is the region where the greater amount of depression has resulted in the invasion of sea waters. The northern edge of the Inner Zone is studded with gigantic lava domes formed by Mount Dai, which, together with volcanic Mount Aso, bury a considerable part of the western extension of the Inland Sea in central Kyushu.

Taishō Pond in Kamikōchi Valley, Chūbu region. Beyond are the peaks of Mount Hotaka, highest mountain in the Hida Range, which is the northernmost range in the Japanese Alps.
T. Okuda—Aspect Picture Library, London

The Ryukyu Islands Region constitutes the main portion of the Ryukyu arc, which penetrates into Kyushu as the West Japan Volcanic Belt and terminates at Mount Aso. The influence of the arc is also seen in the trend of the many elongated islands off western Kyushu, including the Koshiki, Gotō, and Tsushima islands. The islands of the Izu-Ogasawara Region, to the east of the Ryukyu arc, consist of a number of volcanoes on the submarine ridge of the Izu-Marina arc and the Bonin Islands, which include Peel Island and Iwo Jima (Iō-jima).

DRAINAGE AND SOILS

Drainage. Japan's rivers are short and swift-running, supplied by small drainage basins. The most significant rivers are the Teshio and Ishikari of Hokkaido; the Kitakami, Tone, Shinano, Kiso, and Tenryū of Honshu; and the Chikugo of Kyushu. Some of the rivers from the volcanic areas of northeastern Honshu are acidic and are useless for irrigation and other purposes.

Lakes

Lake Biwa, the largest in Japan, covers 260 square miles of central Honshu. All other major lakes are in the northeast. Most of the coastal lakes, such as Lakes Kasumi and Hamana of Honshu, are drowned former valleys, the bay

© 2005 Encyclopædia Britannica, Inc.


Key to Prefectures
(shown by number on map):
1 HOKKAIDO
2 AOMORI
3 AKITA
4 IWATE
5 MIYAGI
6 YAMAGATA
7 NIIGATA
8 FUKUSHIMA
9 IBARAKI
10 TOCHIGI
11 GUMMA
12 SAITAMA
13 TOKYO
14 CHIBA
15 KANAGAWA
16 YAMANASHI
17 SHIZUOKA
18 NAGANO
19 TOYAMA
20 ISHIKAWA
21 FUKUI
22 GIFU
23 AICHI
24 MIE
25 SHIGA
26 KYŌTO
27 NARA
28 WAKAYAMA
29 ŌSAKA
30 HYŌGO
31 TOTTORI
32 OKAYAMA
33 KAGAWA
34 TOKUSHIMA
35 KŌCHI
36 EHIME
37 HIROSHIMA
38 SHIMANE
39 YAMAGUCHI
40 NAGASAKI
41 SAGA
42 FUKUOKA
43 ŌITA
44 KUMAMOTO
45 MIYAZAKI
46 KAGOSHIMA
47 OKINAWA
HOKKAIDO
Hakodate
Tsugaru Strait
Seikan Tunnel
Ō ISLAND
SHIMOKITA PENINSULA
TSUGARU PENINSULA
Mutsu Bay
Aomori
Hirosaki
Hachinohe
HAKKŌDA MTS.
Lake Towada
TOWADA HACHIMANTAI NATIONAL PARK
KITAKAMI RANGE
RIKUCHŪ-KAIGAN NATIONAL PARK
Noshiro
OGA PENINSULA
Akita
Morioka
TŌHOKU
Mount Iwate 2041
Lake Tazawa
Miyako
TOBI ISLAND
Honjō
Yokote
Kamaishi
Mount Chōkai 2236
Sakata
Ōfunato
Shinjō
Tsuruoka
AWA ISLAND
Mogami
Yamagata
Sendai
Ishinomaki
Sendai Bay
OSHIKA PENINSULA
SADO
Mount Kimpoku 1173
Ryōtsu
Aikawa
Murakami
BANDAI-ASAHI NATIONAL PARK
Shibata
Niigata
Fukushima
Yonezawa
OU RANGE
ABUKUMA RANGE
NIIGATA PLAIN
Nagaoka
Aizu-Wakamatsu
Kōriyama
Lake Inawashiro
Shinano
Kashiwazaki
SEA OF JAPAN (EAST SEA)
Wajima
Suzu
Toyama Bay
NOTO PENINSULA
Nanao
Jyoetsu
Mount Myōkō 2446
JŌ-SHIN-ETSU KŌGEN NATIONAL PARK
NIKKŌ NATIONAL PARK
Iwaki
HONSHU
Takaoka
Kanazawa
Toyama
Nagano
Numata
Nikkō
Hitachi
Kita-Ibaraki
Komatsu
CHŪBU SANGAKU NATIONAL PARK
HIDA RANGE
Mount Hotaka 3190
Ueda
Maebashi
Takasaki
Kiryū
KANTŌ
Utsunomiya
Mito
Kashima Bay
Katsuta
Matsumoto
Isesaki
Ashikaga
Oyama
Tsuchiura
Lake Kasumi
Takayama
Lake Suwa
Kumagaya
KANTŌ PLAIN
Fukui
Mount Ontake 3063
Okaya
Ōmiya
Urawa
Koshigaya
Chōshi
OKI ISLANDS
Oki Strait
DAISEN-OKI NATIONAL PARK
SAN'IN-KAIGAN NATIONAL PARK
TANGO PENINSULA
Wakasa Bay
CHŪBU
KISO RANGE
AKAISHI RANGE
Kawaguchi
Funabashi
Ichikawa
Tokyo
Narashino
Chiba
Matsue
Yonago
Tottori
Maizuru
Gifu
Kōfu
Hachiōji
Sagamihara
Machida
Kawasaki
Yokohama
Ichihara
Mobara
BŌSŌ PENINSULA
Yokosuka
Kamogawa
Lake Biwa
Ōgaki
Ichinomiya
Nagoya
Kyoto
Mount Fuji 3776
Fujinomiya
Odawara
Tokyo Bay
CHŪGOKU RANGE
KINKI
Himeji
Kōbe
Ōsaka
Sakai
Nara
Wakayama
Shimizu
Shizuoka
Numazu
Toyota
Okazaki
Toyohashi
Hamamatsu
Lake Hamana
Suruga Bay
Sagami Bay
IZU PENINSULA
Shimoda
Ō ISLAND
Mount Mihara 754
FUJI-HAKONE-IZU NATIONAL PARK
Okayama
Kurashiki
Fukuyama
Tsu
Ise Bay
Matsusaka
Ise
ISE-SHIMA NATIONAL PARK
Owase
SETO-NAIKAI NATIONAL PARK
Takamatsu
Tokushima
Imabari
Matsuyama
SHIKOKU MOUNTAINS
KII MOUNTAINS
YOSHINO-KUMANO NATIONAL PARK
Kii Channel
Kumano
Tanabe
KII PENINSULA
Nachi-Katsuura
Kōchi
Tosa Bay
SHIKOKU
Nakamura
Sukumo
ASHIZURI-UWAKAI NATIONAL PARK
IZU ISLANDS
Mount O 813
AOGA ISLAND
BEYONEISU ROCKS
SUMISU ISLAND
PACIFIC OCEAN
TORI ISLAND
Scale 1:8,652,000
1 inch equals approx. 137 miles
0 25 50 75 100 mi
0 40 80 120 160 km
SŌFU-GAN
MUKOJIMA ISLANDS
NISHINO ISLAND
BONIN ISLANDS
CHICHIJIMA ISLANDS
HAHAJIMA ISLANDS
KITA IWO JIMA
VOLCANO ISLANDS
IWO JIMA
MINAMI IWO JIMA

MAP INDEX

Political subdivisions

Prefectures

Aichi 35 00 N 137 15 E
Akita 40 00 N 140 30 E
Aomori 40 45 N 141 00 E
Chiba 35 30 N 140 20 E
Ehime 33 45 N 132 45 E
Fukui 36 00 N 136 15 E
Fukuoka 33 30 N 130 30 E
Fukushima 37 30 N 140 15 E
Gifu 35 45 N 137 00 E
Gumma 36 30 N 139 00 E
Hiroshima 34 30 N 133 00 E
Hokkaido 44 00 N 143 00 E
Hyōgo 35 00 N 135 00 E
Ibaraki 36 30 N 140 30 E
Ishikawa 36 45 N 136 45 E
Iwate 39 30 N 141 30 E
Kagawa 34 15 N 134 00 E
Kagoshima 30 00 N 130 00 E
Kanagawa 35 30 N 139 30 E
Kōchi 33 40 N 133 30 E
Kumamoto 32 30 N 130 40 E
Mie 34 30 N 136 30 E
Miyagi 38 30 N 141 00 E
Miyazaki 32 15 N 131 30 E
Nagano 36 00 N 138 00 E
Nagasaki 33 00 N 129 30 E
Nara 34 30 N 136 00 E
Niigata 37 30 N 139 00 E
Ōita 33 15 N 131 30 E
Okayama 35 00 N 134 00 E
Okinawa 26 42 N 128 11 E
Saga 33 20 N 130 00 E
Saitama 36 00 N 139 30 E
Shiga 35 15 N 136 00 E
Shimane 35 00 N 132 30 E
Shizuoka 35 00 N 138 00 E
Tochigi 36 45 N 139 45 E
Tokushima 34 15 N 134 00 E
Tottori 35 30 N 134 00 E
Toyama 36 30 N 137 30 E
Wakayama 34 00 N 135 30 E
Yamagata 38 30 N 140 15 E
Yamaguchi 34 20 N 131 30 E
Yamanashi 35 30 N 138 30 E

Urban prefectures

Kyōto 35 10 N 135 30 E
Ōsaka 34 30 N 135 30 E

Metropolis

Tokyo 35 45 N 139 30 E

Cities and towns

Abashiri 44 01 N 144 17 E
Ageo 35 58 N 139 36 E
Aikawa 38 02 N 138 15 E
Aizu-Wakamatsu 37 30 N 139 55 E
Akashi 34 38 N 134 59 E
Akita 39 43 N 140 07 E
Akkeshi 43 02 N 144 51 E
Ako 34 45 N 134 24 E
Akune 32 01 N 130 12 E
Amagasaki 34 43 N 135 25 E
Anan 33 55 N 134 40 E
Anjō 34 57 N 137 05 E
Aomori 40 49 N 140 45 E
Arao 32 59 N 130 29 E
Arida 34 05 N 135 07 E
Arikawa 32 59 N 129 07 E
Asahikawa 43 46 N 142 22 E
Ashibe 33 48 N 129 46 E
Ashibetsu 43 31 N 142 11 E
Ashikaga 36 20 N 139 27 E
Ashiya 33 53 N 130 40 E
Asuka 34 28 N 135 49 E
Beppu 33 17 N 131 30 E
Bibai 43 19 N 141 52 E
Bihoro 43 49 N 144 07 E
Chiba 35 36 N 140 07 E
Chichibu 35 59 N 139 05 E
Chitose 42 49 N 141 39 E
Chōshi 35 44 N 140 50 E
Date 42 27 N 140 51 E
Ebetsu 43 07 N 141 34 E
Engaru 44 03 N 143 31 E
Eniwa 42 53 N 141 34 E
Esashi 41 52 N 140 07 E
Esashi 44 56 N 142 35 E
Fuji 35 12 N 138 43 E
Fujieda 34 52 N 138 16 E
Fujinomiya 35 18 N 138 37 E
Fujisawa 35 21 N 139 29 E
Fukagawa 43 43 N 142 03 E
Fukaya 36 12 N 139 17 E
Fukuchiyama 35 18 N 135 07 E
Fukue 32 41 N 128 50 E
Fukui 36 04 N 136 13 E
Fukuoka 33 35 N 130 24 E
Fukushima 37 45 N 140 28 E
Fukushima 41 29 N 140 15 E
Fukuyama 34 29 N 133 22 E
Funabashi 35 42 N 139 59 E
Furano 43 15 N 142 22 E
Furukawa 38 34 N 140 58 E
Futtsu 35 19 N 139 49 E
Gifu 35 25 N 136 45 E
Ginowan 26 16 N 127 45 E
Gobō 33 53 N 135 10 E
Gōnoura 33 45 N 129 41 E
Goshogawara 40 48 N 140 27 E
Gotemba 35 18 N 138 55 E
Gyōda 36 08 N 139 28 E
Haboro 44 22 N 141 42 E
Hachinohe 40 30 N 141 29 E
Hachiōji 35 39 N 139 20 E
Hagi 34 24 N 131 25 E
Hakata 34 07 N 132 57 E
Hakodate 41 45 N 140 43 E
Hakone 35 11 N 139 02 E
Hakui 36 53 N 136 47 E
Hamada 34 53 N 132 05 E
Hamakita 34 49 N 137 47 E
Hamamatsu 34 42 N 137 44 E
Hanamaki 39 23 N 141 07 E
Higashi-Ōsaka 34 40 N 135 36 E
Higashir Rishiri 45 12 N 141 07 E
Himeji 34 49 N 134 42 E
Hirado 33 22 N 129 33 E
Hirara 24 47 N 125 17 E
Hiratsuka 35 19 N 139 21 E
Hirosaki 40 35 N 140 28 E
Hiroshima 34 24 N 132 27 E
Hita 33 19 N 130 56 E
Hitachi 36 36 N 140 39 E
Hōfu 34 03 N 131 34 E
Hondo 32 27 N 130 12 E
Honjō 39 23 N 140 03 E
Hyūga 32 18 N 131 45 E
Ibusuki 31 13 N 130 36 E
Ichihara 35 25 N 140 07 E
Ichikawa 35 44 N 139 55 E
Ichinomiya 35 18 N 136 48 E
Ichinoseki 38 55 N 141 08 E
Iizuka 33 38 N 130 41 E
Imabari 34 03 N 133 00 E
Imagane 42 25 N 140 01 E
Imari 33 16 N 129 53 E
Ina 35 50 N 137 57 E
Innoshima 34 19 N 133 10 E
Isahaya 32 50 N 130 03 E
Ise 34 29 N 136 42 E
Isesaki 36 19 N 139 12 E
Ishigaki 24 23 N 124 08 E
Ishinomaki 38 25 N 141 18 E
Ishioka 36 11 N 140 16 E
Itō 34 58 N 139 05 E
Iwaki 37 05 N 140 50 E
Iwakuni 34 09 N 132 11 E
Iwamizawa 43 12 N 141 46 E
Iwanai 42 58 N 140 30 E
Izuhara 34 12 N 129 17 E
Izumi 38 21 N 140 48 E
Izumi 32 04 N 130 25 E
Izumi-Sano 34 25 N 135 19 E
Izumo 35 22 N 132 46 E
Jyoetsu 37 07 N 138 15 E
Kaga 36 18 N 136 20 E
Kagoshima 31 36 N 130 33 E
Kainan 34 09 N 135 12 E
Kakamigahara 35 24 N 136 53 E
Kamaishi 39 16 N 141 53 E
Kamiiso 41 49 N 140 39 E
Kamiyaku 30 23 N 130 30 E
Kamogawa 35 06 N 140 06 E
Kanazawa 36 34 N 136 39 E
Kanoya 31 23 N 130 51 E
Kanuma 36 33 N 139 44 E
Karatsu 33 26 N 129 58 E
Kariya 34 59 N 136 59 E
Kasai 34 46 N 137 48 E
Kasaoka 34 30 N 133 30 E
Kashiwazaki 37 22 N 138 33 E
Kasugai 35 15 N 137 00 E
Katsuta 36 24 N 140 33 E
Kawaguchi 35 48 N 139 43 E
Kawasaki 35 32 N 139 43 E
Kesennuma 38 54 N 141 35 E
Kikai 28 20 N 129 50 E
Kiryū 36 24 N 139 20 E
Kishiwada 34 28 N 135 22 E
Kita-Ibaraki 36 50 N 140 40 E
Kita-Kyūshū 33 50 N 130 50 E
Kitakami 39 18 N 141 07 E
Kitami 43 48 N 143 54 E
Kōbe 34 41 N 135 10 E
Kōchi 33 33 N 133 33 E
Kōfu 35 39 N 138 35 E
Koga 36 11 N 139 43 E
Kokubu 31 44 N 130 46 E
Komaki 35 17 N 136 55 E
Komatsu 36 24 N 136 27 E
Komatsushima 34 00 N 134 35 E
Kōriyama 37 24 N 140 23 E
Koshigaya 35 54 N 139 47 E
Kudamatsu 34 00 N 131 52 E
Kuji 40 11 N 141 46 E
Kumagaya 36 08 N 139 23 E
Kumamoto 32 48 N 130 43 E
Kumano 33 57 N 136 05 E
Kurashiki 34 35 N 133 46 E
Kurayoshi 35 26 N 133 49 E
Kure 34 14 N 132 34 E
Kuroiso 36 58 N 140 03 E
Kurume 33 19 N 130 31 E
Kushikino 31 44 N 130 16 E
Kushiro 42 58 N 144 23 E
Kutchan 42 54 N 140 45 E
Kuwana 35 04 N 136 42 E
Kyōto 35 00 N 135 45 E
Machida 35 34 N 139 25 E
Maebashi 36 23 N 139 04 E
Maizuru 35 27 N 135 20 E
Makurazaki 31 16 N 130 19 E
Marugame 34 17 N 133 47 E
Masuda 34 40 N 131 51 E
Matsudo 35 47 N 139 54 E
Matsue 35 28 N 133 04 E
Matsumae 41 26 N 140 07 E
Matsumoto 36 14 N 137 58 E
Matsuto 36 31 N 136 33 E
Matsuyama 33 50 N 132 45 E
Matsuzaka 34 34 N 136 32 E
Mihara 34 24 N 133 05 E
Minamata 32 13 N 130 24 E
Minatomachi 43 12 N 140 46 E
Mine 34 11 N 131 12 E
Mino 35 32 N 136 55 E
Misawa 40 41 N 141 24 E
Mishima 35 07 N 138 55 E
Mito 36 22 N 140 28 E
Miyako 39 38 N 141 57 E
Miyakonojō 31 44 N 131 04 E
Miyazaki 31 52 N 131 25 E
Miyoshi 34 48 N 132 51 E
Mizusawa 39 08 N 141 08 E
Mobara 35 25 N 140 18 E
Mombetsu 44 21 N 143 22 E
Mori 42 06 N 140 35 E
Morioka 39 42 N 141 09 E
Murakami 38 14 N 139 29 E
Muroran 42 18 N 140 59 E
Muroto 33 22 N 134 10 E
Mutsu 41 17 N 141 13 E
Nachi-Katsuura 33 39 N 135 53 E
Nagano 36 39 N 138 11 E
Nagaoka 37 27 N 138 51 E
Nagasaki 32 48 N 129 55 E
Nagato 34 23 N 131 12 E
Nago 26 35 N 127 59 E
Nagoya 35 10 N 136 55 E
Naha 26 13 N 127 40 E
Nakamura 32 59 N 132 56 E
Nakatsu 33 34 N 131 13 E
Nanao 37 03 N 136 58 E
Nankoku 33 35 N 133 39 E
Nara 34 41 N 135 50 E
Narashino 35 41 N 140 03 E
Naruto 34 11 N 134 37 E
Nase 28 23 N 129 30 E
Natori 38 10 N 140 53 E
Nayoro 44 21 N 142 28 E
Nemuro 43 20 N 145 35 E
Nichinan 31 39 N 131 19 E
Niigata 37 55 N 139 03 E
Niihama 33 58 N 133 16 E
Niitsu 37 48 N 139 07 E
Nikkō 36 45 N 139 37 E
Nishinomiya 34 43 N 135 20 E
Nishinoomote 30 44 N 131 00 E
Nishio 34 52 N 137 03 E
Nobeoka 32 35 N 131 40 E
Noboribetsu 42 27 N 141 11 E
Nōgata 33 44 N 130 44 E
Noshiro 40 12 N 140 02 E
Numata 36 38 N 139 03 E
Numazu 35 06 N 138 52 E
Obama 35 30 N 135 45 E
Obihiro 42 55 N 143 12 E
Oda 35 17 N 132 37 E
Ōdate 40 18 N 140 33 E
Odawara 35 15 N 139 10 E
Ōfunato 39 04 N 141 43 E
Ōgaki 35 21 N 136 37 E
Ōhata 41 24 N 141 10 E
Ōita 33 14 N 131 36 E
Ōka 34 06 N 136 13 E
Ōkawa 33 12 N 130 22 E
Okaya 36 03 N 138 03 E
Okayama 34 39 N 133 55 E
Okazaki 34 57 N 137 10 E
Okinawa (Koza) 26 21 N 127 48 E
Ōmiya 35 54 N 139 38 E
Ōmura 32 54 N 129 57 E
Ōmuta 33 02 N 130 27 E
Ōno 35 59 N 136 29 E
Onomichi 34 25 N 133 12 E
Ōsaka 34 40 N 135 30 E
Ōta 36 18 N 139 22 E
Otaru 43 13 N 141 00 E
Otofuke 42 59 N 143 12 E
Ōtsu 35 00 N 135 52 E
Owase 34 02 N 136 11 E
Oyama 36 18 N 139 48 E
Rumoi 43 56 N 141 39 E
Ryōtsu 38 05 N 138 26 E
Sabae 35 57 N 136 11 E
Saga 33 15 N 130 18 E
Sagamihara 35 33 N 139 21 E
Saijo 33 55 N 133 11 E
Saiki 32 58 N 131 53 E
Saito 32 10 N 131 21 E
Sakai 34 35 N 135 28 E
Sakaide 34 19 N 133 54 E
Sakata 38 55 N 139 50 E
Sanjo 37 37 N 138 57 E
Sapporo 43 03 N 141 21 E
Sasebo 33 10 N 129 43 E
Sawara 35 53 N 140 30 E
Sendai 31 49 N 130 18 E
Sendai 38 15 N 140 53 E
Seto 35 14 N 137 06 E
Setouchi 28 08 N 129 15 E
Shari 43 55 N 144 40 E
Shibata 37 57 N 139 20 E
Shibetsu 43 40 N 145 08 E
Shibetsu 44 10 N 142 23 E
Shimada 34 49 N 138 11 E
Shimizu 35 01 N 138 29 E
Shimoda 34 40 N 138 57 E
Shimonoseki 33 57 N 130 57 E
Shinjō 38 46 N 140 18 E
Shiogama 38 19 N 141 01 E
Shirakawa 37 07 N 140 13 E
Shiranuka 42 57 N 144 05 E
Shiroishi 38 00 N 140 37 E
Shizunai 42 20 N 142 22 E
Shizuoka 34 58 N 138 23 E
Sukagawa 37 17 N 140 23 E
Sukumo 32 56 N 132 44 E
Susaki 33 22 N 133 17 E
Suwa 36 02 N 138 08 E
Suzu 37 25 N 137 15 E
Suzuka 34 52 N 136 32 E
Takahashi 34 40 N 133 35 E
Takamatsu 34 20 N 134 03 E
Takaoka 36 45 N 137 01 E
Takasaki 36 20 N 139 01 E
Takatsuki 34 51 N 135 37 E
Takayama 36 08 N 137 15 E
Takefu 35 54 N 136 10 E
Takikawa 43 33 N 141 54 E
Tamano 34 28 N 133 56 E
Tanabe 33 44 N 135 22 E
Tateyama 34 59 N 139 52 E
Tendō 38 21 N 140 22 E
Tōkamachi 37 08 N 138 46 E
Tokunoshima 27 48 N 128 59 E
Tokushima 34 04 N 134 34 E
Tokuyama 34 03 N 131 49 E
Tokyo 35 42 N 139 46 E
Tomakomai 42 38 N 141 36 E
Tosa 34 26 N 135 48 E

Tosu 33 22 N 130 31 E
Tottori 35 30 N 134 14 E
Towada 40 18 N 140 52 E
Toyama 36 41 N 137 13 E
Tōyō 33 22 N 134 18 E
Toyohashi 34 46 N 137 23 E
Toyokawa 34 49 N 137 24 E
Toyooka 35 32 N 134 50 E
Toyota 35 05 N 137 10 E
Tsu 34 43 N 136 31 E
Tsuchiura 36 05 N 140 12 E
Tsukumi 33 04 N 131 52 E
Tsuruga 35 39 N 136 04 E
Tsuruoka 38 44 N 139 50 E
Tsuyama 35 03 N 134 00 E
Ube 33 56 N 131 15 E
Ueda 36 24 N 138 16 E
Ueno 34 45 N 136 08 E
Uji 34 53 N 135 48 E
Uozu 36 48 N 137 24 E
Urakawa 42 09 N 142 47 E
Urasoe 26 15 N 127 43 E
Urawa 35 51 N 139 39 E
Usa 33 27 N 133 27 E
Ushibuka 32 11 N 130 01 E
Usuki 33 08 N 131 49 E
Utsunomiya 36 33 N 139 52 E
Uwajima 33 13 N 132 34 E
Wajima 37 24 N 136 54 E
Wakayama 34 13 N 135 11 E
Wakkanai 45 25 N 141 40 E
Yaizu 34 52 N 138 20 E
Yamagata 38 15 N 140 20 E
Yamaguchi 34 10 N 131 29 E
Yao 34 37 N 135 36 E
Yasugi 35 26 N 133 15 E
Yatsushiro 32 30 N 130 36 E
Yawata 34 52 N 135 42 E
Yawatahama 33 27 N 132 25 E
Yokkaichi 34 58 N 136 37 E
Yokohama 35 27 N 139 39 E
Yokosuka 35 18 N 139 40 E
Yokote 39 18 N 140 34 E
Yonago 35 26 N 133 20 E
Yonezawa 37 55 N 140 07 E
Yūbari 43 04 N 141 59 E

Physical features and points of interest

Abashiri Bay 44 00 N 144 30 E
Abukuma, *river* 37 50 N 140 52 E
Abukuma Range 37 30 N 140 45 E
Agano, *river* 37 57 N 139 08 E
Aguni Island 26 35 N 127 13 E
Akaishi Range 35 25 N 138 10 E
Akan NP 43 30 N 144 15 E
Amakusa Islands 32 20 N 130 05 E
Amami Great Island 28 15 N 129 20 E
Amami Islands 28 16 N 129 21 E
Aoga Island 32 28 N 139 46 E
Ariake Sea 33 05 N 130 15 E
Asahi, *river* 35 00 N 133 58 E
Asahi, Mount, *volcano* 43 40 N 142 51 E
Ashizuri-Uwakai NP 33 00 N 132 30 E
Aso, Mount, *volcano* 32 53 N 131 06 E
Aso-Kuju NP 33 00 N 131 07 E
Awa Island 38 27 N 139 15 E
Awaji Island 34 20 N 134 50 E
Bandai-Asahi NP 38 16 N 139 57 E
Beyoneisu Rocks 31 55 N 139 55 E
Biwa, Lake 35 15 N 136 05 E
Bonin Islands 27 00 N 142 10 E
Bōsō Peninsula 35 20 N 140 10 E
Chichibu-Tama NP 35 52 N 139 00 E
Chichijima Islands 27 06 N 142 12 E
Chikugo, *river* 33 09 N 130 21 E
Chōkai, Mount 39 06 N 140 03 E
Chūbu Sangaku NP 36 30 N 137 41 E
Chūgoku Range 35 15 N 134 00 E
Dai, Mount, *volcano* 35 22 N 133 32 E
Daisen-Oki NP 35 33 N 133 10 E
Daisetsu Mountains 43 41 N 142 53 E
Daisetsuzan NP 43 30 N 142 57 E
Danjo Archipelago, *islands* 32 02 N 128 22 E
Erimo, Cape 41 55 N 143 15 E
Fuji, Mount, *volcano* 35 22 N 138 44 E
Fuji-Hakone-Izu NP 35 21 N 138 44 E
Gonō, *river* 35 02 N 132 13 E
Gotō Islands 32 50 N 129 00 E
Hahajima Islands 26 37 N 142 10 E
Hakkōda Mountains 40 40 N 140 53 E
Hamana, Lake 34 45 N 137 34 E
Hibiki Bay 34 00 N 130 30 E
Hida Range 36 45 N 137 45 E
Hidaka Range 42 35 N 142 45 E
Hokkaido, *island* 44 00 N 143 00 E
Honshu, *island* 36 00 N 138 00 E
Hotaka, Mount 36 17 N 137 39 E
Iheya Island 27 04 N 127 58 E
Iki, *island* 33 47 N 129 43 E
Inawashiro, Lake 37 29 N 140 06 E
Inland Sea 34 20 N 133 30 E
Iriomote Island 24 20 N 123 50 E
Iriomote NP 24 20 N 123 48 E
Ise Bay 34 43 N 136 43 E
Ise-Shima NP 34 23 N 136 48 E
Ishigaki Island 24 24 N 124 12 E
Ishikari, *river* 43 15 N 141 23 E
Ishikari Bay 43 25 N 141 01 E
Ishikari Plain 43 00 N 141 40 E
Iwate, Mount 39 51 N 141 00 E
Iwo Jima, *island* 24 47 N 141 20 E
Iyo Bay 33 30 N 132 15 E
Izu Islands 34 20 N 139 40 E
Izu Peninsula 34 45 N 139 00 E
Japanese Alps, *mountains* 36 00 N 137 00 E
Jō-Shin'etsu Kōgen NP 36 46 N 138 40 E
Kagoshima Bay 31 25 N 130 38 E
Kamui, Cape 43 20 N 140 21 E
Kantō Plain 36 00 N 139 30 E
Kantō Range 35 50 N 138 50 E
Kashima Bay 36 00 N 140 45 E
Kasumi, Lake 36 00 N 140 25 E
Kerama Islands 26 05 N 127 15 E
Kii Channel 34 00 N 134 55 E
Kii Mountains 34 00 N 135 50 E
Kii Peninsula 34 00 N 135 45 E
Kikai Island 28 19 N 129 59 E
Kimpoku, Mount 38 06 N 138 21 E
Kirishima-Yaku NP 31 55 N 130 55 E
Kiso, *river* 35 02 N 137 30 E
Kiso Range 35 40 N 137 45 E
Kita Iwo Jima, *island* 25 26 N 141 17 E
Kitakami, *river* 38 40 N 141 19 E
Kitakami Range 39 30 N 141 30 E
Kitami Mountains 44 22 N 142 43 E
Kōbi Island 25 56 N 123 41 E
Koshiki Islands 31 46 N 129 48 E
Koza, *river* 33 31 N 135 50 E
Kume Island 26 21 N 126 46 E
Kuro Island 30 50 N 129 57 E
Kushiro-Shitsugen NP 43 05 N 145 30 E
Kutcharo, Lake 43 38 N 144 21 E
Kyushu, *island* 33 00 N 131 00 E
Kyushu Mountains 32 35 N 131 05 E
Mi Island 34 41 N 131 42 E
Mihara, Mount 34 43 N 139 23 E
Minami Alps NP 35 45 N 138 15 E
Minami Iwo Jima, *island* 24 14 N 141 28 E
Miyako Island 24 47 N 125 20 E
Miyanoura, Mount 30 20 N 130 31 E
Mogami, *river* 38 55 N 139 48 E
Mukojima Islands 27 37 N 142 10 E
Mutsu Bay 41 00 N 141 00 E
Myōkō, Mount 36 52 N 138 07 E
Naka, *river* 36 20 N 140 36 E
Nantai, Mount 36 43 N 140 25 E
Niigata Plain 37 45 N 139 00 E
Nikkō NP 36 49 N 139 33 E
Nishino Island 27 15 N 140 52 E
Nōbi Plains 35 15 N 137 30 E
Noheji Ogawara, Lake 40 47 N 141 20 E
Noto Peninsula 37 20 N 137 00 E
O, Mount 34 08 N 139 35 E
Ō Island 34 44 N 139 24 E
Oga Peninsula 39 55 N 139 50 E
Oki Islands 36 15 N 133 15 E
Okinawa, *island* 26 42 N 128 11 E
Ōkino-erabu Island 27 22 N 128 35 E
Okushiri Island 42 10 N 139 28 E
Ōnaruto Bridge 34 13 N 134 39 E
Ontake, Mount, *volcano* 35 53 N 137 39 E
Ōsaka Bay 34 30 N 135 18 E
Oshika Peninsula 38 20 N 141 30 E
Oshima Peninsula 42 00 N 140 30 E
Ōsumi Islands 30 30 N 130 00 E
Ōsumi Peninsula 31 20 N 130 55 E
Ōu Range 39 00 N 140 45 E
Rebun Island 45 22 N 141 02 E
Rikuchu-Kaigan NP 39 30 N 142 00 E
Rishiri Island 45 11 N 141 15 E
Rishiri-Rebun-Sarobetsu NP 45 11 N 141 15 E
Ryukyu Islands 29 00 N 129 00 E
Sado, *island* 38 00 N 138 25 E
Sagami Bay 35 15 N 139 25 E
Saikai NP 33 05 N 129 38 E
Sakishima Islands 24 46 N 124 00 E
San'in-Kaigan NP 35 40 N 134 50 E
Saroma, Lake 44 08 N 143 50 E
Satsuma Peninsula 31 25 N 130 25 E
Seikan Tunnel 41 17 N 140 49 E
Sekibi Island 24 55 N 124 34 E
Sendai Bay 38 10 N 141 20 E
Sennokura, Mount 36 49 N 138 50 E
Seto Bridge 34 23 N 133 40 E
Seto-Naikai NP 34 30 N 134 15 E
Shari, Mount 43 46 N 144 43 E
Shikoku, *island* 33 45 N 133 30 E
Shikoku Mountains 33 50 N 133 26 E
Shikotsu, Lake 42 45 N 141 20 E
Shikotsu-Tōya NP 42 47 N 141 10 E
Shimokita Peninsula 41 20 N 141 00 E
Shinano, *river* 37 00 N 138 45 E
Shiretoko, Cape 44 21 N 145 20 E
Shiretoko NP 44 21 N 145 20 E
Shokambetsu, Mount 43 43 N 141 31 E
Shōwa Volcano 42 32 N 140 52 E
Sōfu-Gan, *island* 29 49 N 140 21 E
Sōya, Cape 45 31 N 141 56 E
Sumisu Island 31 27 N 140 03 E
Suō Sea 33 50 N 131 30 E
Suruga Bay 34 51 N 138 33 E
Suwa, Lake 36 03 N 138 05 E
Tanega Island 30 40 N 131 00 E
Tango Peninsula 35 40 N 135 10 E
Tazawa, Lake 39 43 N 140 40 E
Tenryū, *river* 34 39 N 137 48 E
Teshio, *river* 44 40 N 142 20 E
Teshio Range 44 15 N 142 05 E
Tobi Island 39 12 N 139 34 E
Tokachi, *river* 42 44 N 143 42 E
Tokachi, Mount 43 25 N 142 41 E
Tokachi Plain 43 00 N 143 30 E
Tokara Islands 29 37 N 129 43 E
Tokoro, *river* 43 40 N 143 45 E
Tokuno Island 27 45 N 128 58 E
Tokyo Bay 35 25 N 139 47 E
Tone, *river* 35 44 N 140 51 E
Tori Island 30 29 N 140 19 E
Tosa Bay 33 20 N 133 40 E
Towada, Lake 40 28 N 140 55 E
Towada-Hachimantai NP 40 35 N 140 53 E
Tōya, Lake 42 35 N 140 51 E
Toyama Bay 36 50 N 137 10 E
Tsugaru Peninsula 36 50 N 137 10 E
Tsushima, Islands 34 20 N 129 20 E
Uchiura Bay 42 20 N 140 40 E
Unzen, Mount, *volcano* 32 45 N 130 17 E
Unzen-Amakusa NP 32 45 N 130 17 E
Uoturi Island 25 45 N 123 29 E
Volcano Islands 25 00 N 141 00 E
Wakasa Bay 35 45 N 135 40 E
Yaku Island 30 20 N 130 30 E
Yodo, *river* 34 41 N 135 31 E
Yonaguni Island 24 27 N 123 00 E
Yoron Island 27 02 N 128 26 E
Yoshino-Kumano NP 34 00 N 135 56 E

NP = National Park

mouths of which have been dammed by sandbars. Inland lakes such as Biwa, Suwa, and Inawashiro of Honshu occupy tectonic depressions of geologically recent fault origin. Lakes of volcanic origin (*e.g.,* Kutcharo of Hokkaido and Towada and Ashi of Honshu) outnumber all other types.

Soils. The soils of Japan are customarily divided from northeast to southwest into a weak podzolic (soils with a thin organic mineral layer over a gray leached layer) zone, a brown-earth zone, and a red-earth zone. There are some local variations. The northern half of the Tōhoku area of northern Honshu is included in the area of brown forest soils. The northern tip of Hokkaido is classed as a subzone of the podzolic soils; the remainder of the island is included in the subzone of the acidic brown forest soils. Most of western Honshu is a transitional zone. Yellow-brown forest soils extend along the Pacific coast from southern Tōhoku to southern Kyushu, while red and yellow soils are confined to the Ryukyu Islands. The widespread reddish soils are the products of a former warmer, more humid climate. Immature volcanic ash soils occur on the uplands. *Kuroboku* soils (black soils rich in humus) are found on terraces, hills, and gentle slopes throughout Japan, while gley (sticky, blue-gray compact) soils are found in the poorly drained lowlands. Peat soils occupy the moors in Hokkaido and Tōhoku. Muck (dark soil, containing a high percentage of organic matter) and gley paddy soils are the products of years of rice culture. Polder soils (those reclaimed from the sea) are widely distributed. Soil fertility increases in the lowlands where agriculture is practiced, the result of a combination of alluvium washed down from the uplands and centuries of intense reworking of the soil by rice farmers.

CLIMATE

In general, Japan's climate is characterized as monsoonal (*i.e.,* governed by seasonal winds). The main influences are the country's latitudinal extent, the surrounding oceans, and its proximity to the neighbouring Asian landmass. There are numerous local climatic variations, the result of relief features. In winter, the high pressure zone over eastern Siberia and the low pressure zone over the western Pacific result in an eastward flow of cold air (the winter monsoon) from late September to late March that picks up moisture over the Sea of Japan. The winter monsoon deposits its moisture as rain or snow on the side of Japan facing the Sea of Japan and brings dry, windy weather to the Pacific side. The pressure systems are reversed during the summer, and air movements from the east and south (the summer monsoon) from mid-April to early September bring warmer temperatures and rain. Cyclonic storms and frequent and destructive typhoons occur during late summer and early fall, especially in the southwest.

The effects of ocean currents

The warm waters of the Kuroshio (Japan Current) flow northward along Japan's Pacific coast as far as latitude 35° N. The Tsushima Current branches westward from the Kuroshio off southern Kyushu and washes the coasts of Honshu and Hokkaido along the Sea of Japan; it is this current that lends moisture to the winter monsoon. The counterpart of the Labrador Current, the cold Oya (Kuril) Current, flows southeastward from the Bering Sea along the east coast of Hokkaido and northeastern Honshu. Its waters meet those of the Kuroshio, causing dense sea fogs in summer, especially off Hokkaido.

The physical feature that most affects climate is the mountainous backbone of the islands. The ranges interrupt the monsoonal winds and cause the gloomy weather and heavy snows of winter along the Sea of Japan coast and the bright and windy winter weather along the Pacific. Temperatures and annual precipitation are about the same on both coasts, but they drop noticeably in the mountainous interior.

Temperature. Temperatures are generally warmer in the south than in the north, and the transitional seasons of spring and fall are shorter in the north. At Asahikawa, in central Hokkaido, the mean temperature in January, the coldest month, is 16° F (−9° C), and the mean temperature in August, the hottest month, is 70° F (21° C), with an annual average temperature of 43° F (6° C). At Tokyo, the mean temperature for January is 39° F (4° C), the mean for August 81° F (27° C), and the annual average 59° F (15° C). Inland from Tokyo, Matsumoto is cooler, with an annual average temperature of 52° F (11° C), whereas an annual average of 57° F (14° C) occurs on the Sea of Japan coast at Kanazawa. The warmest temperatures occur on Kyushu and the southern islands; at Kagoshima, the mean temperature for January is 45° F (7° C), the mean for August is 81° F (27° C), and the average is 63° F (17° C).

Precipitation. Precipitation in the form of rain and snow is plentiful throughout the islands. Maximum precipitation falls in the early summer, and the minimum occurs in winter—except on the Sea of Japan coast, which receives the country's highest snowfall. The summer rainy season occurs through June and July; it is known as the *baiu* ("plum rain") because it begins when the plums ripen. Torrential rains accompany the typhoons.

Precipitation patterns vary with topography, but most of the country receives more than 40 inches (1,020 millimetres) annually, mainly as rain during the summer. The smallest amount of precipitation occurs on eastern Hokkaido, where only 37 inches fall annually at Obihiro, whereas the mountainous interior of the Kii Peninsula of central Honshu receives more than 160 inches annually. Varying amounts of snow fall on Japan. From November to April snow blankets Hokkaido, northern and interior Honshu, and the northwest coast.

PLANT AND ANIMAL LIFE

Flora. Much of the original vegetation has been replaced by agriculture or by the introduction of foreign species to the islands. Semitropical rainforest prevails in the Ryukyu and Bonin archipelagoes and contains various kinds of mulberries, camphor, oaks, and ferns (including tree ferns); madder and lianas are found as undergrowth. In the Amami Islands this type of plant life occurs only on lowlands, but it grows at higher altitudes to the south. There are a few mangrove swamps along the southern coast of Kyushu.

Vegetational regions

The laurel forest zone of evergreen, broad-leaved trees extends from the southwestern islands northward to the lowlands of northern Honshu. Camphor, pasania, Japanese evergreen oak, camellia, and holly are typical trees, and various kinds of ferns grow as undergrowth. In Kyushu, the evergreen zone reaches elevations of more than 3,300 feet, but its vertical limit decreases northeastward across Honshu. In general, camphor dominates in the littoral lowlands, pasania in sunny and well-drained sites, and Japanese evergreen oak in the foggy and cloudy inlands. In the southwestern Hondo region (western Honshu, Shikoku, and Kyushu) are ficus and fan palm. The coastal dunes are dominated by pine trees. Natural stands of Japanese cedar occur above 2,300 feet on Yaku Island, south of Kyushu.

Deciduous broad-leaved forests develop in the higher and more northerly portions of the laurel forest zone. In Kyushu, this type of forest occurs above 3,300 feet, but it gradually descends northward to sea level in northern Honshu. Its upper limit reaches 6,000 feet in Shikoku and 5,000 feet in central Honshu. Representative trees are beech, katsura tree, maple, oak, and birch, while various kinds of bamboo grasses grow as undergrowth. All these trees, but especially the maples, are admired for their beautiful fall colours. The trees have been occasionally replaced by larch, false cypress, false arborvitae, Japanese cedar, Japanese red pine, Japanese black pine, and other coniferous species. The deciduous zone extends into western Hokkaido, where beeches terminate at the southwestern peninsula and further northeastward are replaced by basswood and maple.

Coniferous trees are numerous in the north and eastern periphery of Hokkaido up to elevations of 2,300 feet. Sakhalin spruce, Sakhalin fir, blue fir, and Yezo spruce are mixed with such deciduous trees as birch, oak, and maple and dense undergrowth of mosses and lichens. Coniferous trees are mixed with deciduous vegetation in southwestern Hokkaido and occur in the higher portion of central Honshu and Shikoku. High-altitude small shrubs, the creeping pine, and alpine plants grow in the high mountain knots of central Honshu above 8,000 feet. This zone gradually descends northward to the Hakkōda Mountains, in northern Honshu, at 4,600 feet and to the Daisetsu Mountains, in central Hokkaido, at about 3,600 feet.

The cherry tree, long celebrated for its spring blossoms and as one of the symbols of Japan, is planted throughout the country. Many varieties have been cultivated, and natural stands are found in the mountains.

Fauna. Despite the country's large human population, the land mammals of Japan are relatively numerous in the remote, heavily forested mountain regions. These animals include bears, wild boars, raccoon dogs (*tanuki*), foxes, deer, antelope, hares, and weasels; some species are distinct from those of the neighbouring Asian continent. Wild monkeys (the Japanese macaque) inhabit many places; those found at the northern tip of Honshu represent the northern limit of monkey habitation in the world.

Monkeys

Reptiles include sea turtles, freshwater tortoises, sea snakes, and lizards. There are two species of poisonous snakes, but most of the snakes, including the five-foot-long Japanese rat snake, are harmless. Toads, frogs, and newts are common, and the endemic Japanese giant salamander of Kyushu and western Honshu can attain a length of four feet or more. Insect life is typical of a temperate-humid climate; several species have seasonal associations in literature and popular culture, such as the cicada and dragonfly (summer) and the cricket (autumn).

The Japanese archipelago constitutes a major East Asian flyway, and some 600 bird species are either resident or transitory. Water birds are abundant and include gulls, auks, grebes, albatrosses, shearwaters, herons, ducks, geese, swans, and cranes. The cormorant is sometimes trained to catch fish. There are about 150 species of songbirds, as well as eagles, hawks, falcons, pheasant, ptarmigan, quail, owls, and woodpeckers.

The confluence of cold and warm ocean currents near Japan has produced a rich sea life. Japanese waters are inhabited by whales, dolphins, porpoises, and fish such as salmon, sardines, sea bream, mackerel, tuna, trout, herring, gray mullet, smelts, and cod. Crustaceans and mollusks include crabs, shrimp, prawns, clams, and oysters. The rivers and lakes abound in trout, salmon, and crayfish.

SETTLEMENT PATTERNS

Traditional regions. The concept of regions in Japan is inseparable from the historic development of administrative units. Care was always taken to include various physical features in the larger administrative units so as to create a well-balanced geographic whole. Many of the ancient terms for administrative units have survived in the form of place-names.

The Taika reforms of the 7th century established the *ri* (roughly corresponding to the later village community) as the basic social and economic unit and the *gun* (district) as the smallest political unit to be governed by the central government. The *gun* were grouped to form more than 60 *kuni* (provinces), the largest political units, which were ruled by governors appointed by the central government. Several adjacent *kuni* that were linked by a trunk road or a convenient sea route were grouped into a *dō*, the term signifying both the route and the region. The core region of the country was called the Kinai, or the land adjacent to the shifting imperial capitals.

Kantō and Kansai

During the Nara (710–784) and Heian (794–1185) periods, the region of Honshu to the east of the three great mountain barriers of Arachi, Fuwa, and Suzuka north, east, and southeast of Lake Biwa was called Kantō and that to the west Kansai (*kan*, "barrier"; *tō*, "east"; *sai*, "west"). As the empire's frontier shifted to the northeast, Kantō came to signify the region to the east of Hakone barrier, and Kansai gradually came to include limited areas near the capital of Kyōto as far as Ōsaka and present-day Kōbe. Northern areas that had not come under direct control of the central government were called Ezochi (or Yezochi), "Land of the Ezo (Ainu)."

A third regional system was applied after the 10th century, in which *kuni* were amalgamated according to their distance from Kyōto. The larger units were *kingoku*, or proximate *kuni; chūgoku*, or intermediate *kuni;* and *engoku*, or remote *kuni.* Mutsu and Dewa in northeastern Honshu and islands such as Sado, Oki, Tsushima, and Iki were termed *henkyō*, or peripheral, lands.

In 1871 the feudal system was dissolved and the *ken*, or prefectural, system was established. At first the more than 300 prefectures were mostly the former fiefs of feudal lords, who were appointed as governors. Through amalgamation and partition there were frequent changes in the *ken* pattern, until by 1888 the present configuration of 43 *ken* (including Okinawa), three *fu* (urban prefectures) of Tokyo, Ōsaka, and Kōbe, and one *dō* (Hokkaido) was established; in 1943 Tokyo was given the status of *to*, or metropolis.

Early in the 20th century it was recognized that larger geographic divisions were needed. By 1905 a system of eight *chihō* (regions) had been set up that divided the country from northeast to southwest. The *chihō* are Hokkaido, Tōhoku (northern Honshu), Kantō (eastern Honshu), Chūbu (central Honshu), Kinki (west-central Honshu), Chūgoku (western Honshu), Shikoku, and Kyushu (including the Ryukyus). Another system used by some governmental agencies is a modification of the *chihō* system. The Chūbu region, for example, is subdivided into Hokuriku, Tōsan, and Tōkai. In addition, planners have come to refer to the string of industrialized and urbanized areas along the Pacific seaboard between Kantō and northern Kyushu as the Pacific Belt Zone (Taihei-yō Beruto Chitai). This zone includes most of the Japanese cities with populations of more than one million, as well as more than half of the country's total population.

Rural settlement. From the late 19th century, economic and social changes affected even the most remote rural villages, but many traditional aspects have survived. In the villages, many features that are in common with those of other Asian villages are well preserved. Autonomous and cooperative systems of agricultural practices and rituals, as well as mutual assistance among the villagers, have been handed down to the present. These traditions are mixed with modernized farming practices and employment diversification. An autonomous rural unit, generally known as a *mura*, consists of some 30 to 50 or more households. Now called an *aza*, this unit should not be confused with the administrative terms *mura* or *son* in use after 1888.

Considerable local difference is evident in the settlement pattern. Some villages are agglomerated, as are those of the Kinki region; some are dispersed, as in northeastern Shikoku; some are elongated, such as those on the rows of sand dunes in the Niigata Plain and on the natural levees of deltas; while others are scattered on the steeper mountain slopes. These differences are only superficial, although the traditional ties that bound the inhabitants together to form a firm village community are changing as industry moves into the countryside and offers farmers attractive employment options.

Changing status of villages

No village is regarded as purely rural. Those that are near industrialized urban centres include large numbers of commuters and industrial workers. The more remote settlements send out seasonal labourers during the winter months, though outright migration to urban centres is now more common. The villages of Hokkaido are based on commercial agriculture, and each household has direct contact with a nearby town.

Fishing villages were absent in Tōhoku until the beginning of the 17th century, when northward movement began. They originally were dependent upon nearby rice-producing villages, although some dried, salted, or smoked fish found more distant markets. The fishing villages are most numerous in the southwest, where an exchange economy has long been in practice. Mountain villages that depend solely on local products other than rice are exceedingly rare. Many of them were founded after the 17th century, when lumber, charcoal, and other such commedities found markets in the growing towns on the plains. There were also some villages in the mountainous interior of western Tōhoku that relied purely upon hunting, but these have all but disappeared.

Urban settlement. Urbanization is generally of recent origin. Except for the former capital cities of Nara, Kyōto, and Kamakura, no sizable town of any significance appeared before the 16th century. Most of the provincial capitals, or *koku-fu*, of ancient Japan were only administrative centres that contained official residences and were not developed towns. After the latter part of the 16th century, influential temples and feudal lords began to build towns by gathering merchants and craftsmen close to their headquarters. The power of the feudal lords stabilized when they built *jōka-machi* (castle towns), which were located so as to command and control the main transportation routes and surrounding areas; the majority of Japan's important cities, including Tokyo, developed from them.

Next in importance were the port towns, such as Hakata and Sakai. In addition, some of the religious towns eventually grew to a considerable size, as in the case of Ise and Izumo. Under the regime of the Tokugawa shogunate (1603–1867), peaceful conditions fostered nationwide pilgrimages, and temple and shrine towns such as Kyōto and Nara flourished.

Widespread urban growth began in the late 19th century with the development of the international ports of Kōbe, Yokohama, Niigata, Hakodate, and Nagasaki and the naval bases of Yokosuka, Kure, and Sasebo. With industrialization came the rapid growth of Japanese cities, and some of the industrial towns (*e.g.*, Yawata, Niihama, Kawasaki, and Amagasaki) were founded in response to economic development. Most of the former castle towns, and especially those along the Pacific belt, have been expanded directly or indirectly by industrialization. In Hokkaido and in southern Kyushu raw materials and power resources have attracted a limited number of industrial plants, which alone are responsible for the existence of cities such as Tomakomai, Muroran, Nobeoka, and Minamata.

Urban land use

Japanese cities are bewildering mixtures of old and new, East and West. Mixed land use, including agricultural activity, can be found side by side with the most modernized business centres and industrial establishments; and the fragmented, patchwork pattern of land ownership is a formidable obstacle in ever-expanding cities of skyscrapers, subways, and underground plazas. Other serious problems are the shortage of better housing, the increasing use of the automobile, overcrowded public transportation systems, the shortage of open space for recreation, environmental pollution, and the constant menace of earthquakes and floods.

The people

ETHNIC AND LINGUISTIC COMPOSITION

Japanese ethnicity. The Japanese people are members of the Asiatic geographic race and are closely akin to the other peoples of eastern Asia; they constitute the overwhelming majority of the population. During the Tokugawa period, there was a social division of the populace into four classes (warrior, farmer, craftsman, and merchant), with a peer class above and an outcast class below. With the exception of the *burakumin,* the descendants of the former outcast class, this social-class system has almost disappeared. The *burakumin,* however, are still subject to varying degrees of discrimination.

Insofar as a social-class system does persist it does not have the ethnic basis that can exist in multiracial societies, since the Japanese regard themselves as belonging to a single ethnic group. The few exceptions include those classified as resident aliens (particularly Koreans) and Japanese citizens of Ainu and, to a lesser degree, Okinawan origin. Japan also has a small population of Chinese descent.

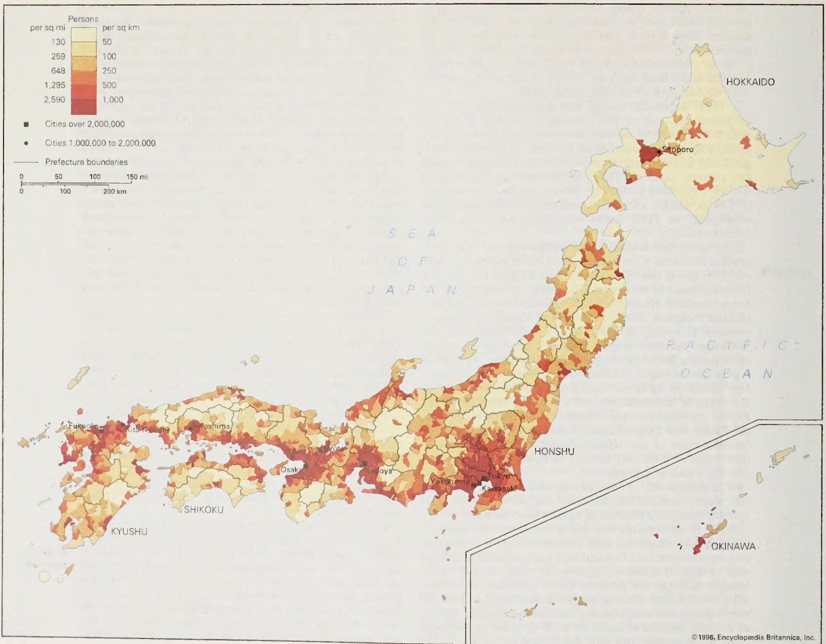
Hundreds of thousands of Koreans migrated to Japan (a great many against their will) before and during World War II, when Korea was a Japanese colony; those remaining after the war and their descendants, the latter born and raised in Japan, do not have Japanese citizenship and face considerable discrimination. Both Ainu and Okinawans are often relegated to a second-class status. The indigenous Ainu largely were assimilated into the general population centuries ago; a few small, scattered groups, however, have maintained their identity in Hokkaido. Before the war there was a tendency to distinguish the people of Okinawa from other Japanese because of perceived differences; this tendency has diminished but not disappeared.

Languages. Japanese is the national language, and Ainu is almost extinct. The Japanese language is generally included in the Altaic linguistic group and is especially akin to Korean, although the vocabularies differ. The introduction of the Chinese writing system and of Chinese literature about the 4th century AD enriched the Japanese vocabulary. Until that time Japanese had no written form, and at first Chinese characters (called *kanji* in Japanese) were used to write Japanese; by the 9th century two syllabaries, known collectively as *kana* (*katakana* and *hiragana*), were developed from them. Since then, a combination of *kanji* and *kana* has been used for written Japanese. Although some 3,000 to 5,000 *kanji* are in general use, the number of characters necessary for a basic vocabulary has been reduced to about 2,000, and the writing of these characters was simplified. Tens of thousands of Western loanwords, principally from English, also have been adopted.

Japanese writing system

The distribution of Japanese nearly coincides with the territory of Japan. Standard Japanese, based on the dialect spoken in Tokyo, was established in the late 19th century through the creation of a national educational system and through more widespread communication. There are many local dialects, which are often mutually unintelligible, but standard Japanese is understood nationwide.

Japanese is broadly divided linguistically into the two



Population density of Japan.

major dialects of Hondo and Nantō. The Hondo dialect is used throughout Japan and may be divided into three major subdialects: Eastern, Western, and Kyushu. The Eastern subdialects were established in the 7th and 8th centuries. After the 17th century there was a vigorous influx of the Kamigata (Kinai) dialect, which was the foundation of standard Japanese. Among the Western dialects, the Kinki dialect was long the standard language of Japan, although the present Kamigata dialect of the Kyōto-Ōsaka region is of recent origin.

The Kyushu dialects have been placed outside the mainstream of linguistic change of the Western dialects and retain some of the 16th-century forms of the latter. They extend as far south as Tanega and Yaku islands. The Nantō dialects are used by Okinawa islanders from the Amami Islands in Kagoshima prefecture to Yonaguni Island at the western end of the archipelago.

RELIGIONS

In Japan, the indigenous religion, Shintō, various sects of Buddhism, and Christianity exist together with some ancient shamanistic practices and a number of "new religions" (*shinkō shukyō*) that have emerged since the 19th century. Not one of the religions is dominant, and each is affected by the others. Thus, it is typical for one person or family to believe in several Shintō gods and at the same time belong to a Buddhist sect. Intense religious feelings are generally lacking except among the adherents of some of the new religions. On the other hand, many Japanese homes contain a Buddhist altar (*butsudan*), at which various rituals—some on a daily basis—commemorate deceased family members.

Shintō is a polytheistic religion. People, commonly major historical figures, as well as natural objects have been enshrined as gods. Some of the Hindu gods and Chinese spirits were also introduced and Japanized. Each rural settlement has at least one shrine of its own, and there are several shrines of national significance, the most important of which is the Grand Shrine of Ise in Mie prefecture.

D.E. Cox—TSW-CLICK/Chicago

Purification shrine in the Kiyomizu temple, Kyōto.

After the Meiji Restoration (1868), Shintō was restructured as a state-supported religion, but this institution was abolished after World War II.

Buddhism

Buddhism, which claims the largest number of adherents after Shintō, was officially introduced into the imperial court from Korea in the mid-6th century AD. Direct contact with central China was maintained, and several sects were introduced, including the Tendai and Shingon sects in the 8th and 9th centuries. These sects have continued to exert profound influence in some parts of Japan. Zen Buddhism, the development of which dates to the late 12th century, has maintained a large following. Most of the major Buddhist sects of modern Japan, however, have descended from those that were modified in the 13th century by monks such as Shinran, who established an offshoot of Pure Land (Jōdo) Buddhism called the True Pure Land sect (Jōdo Shinshū), and Nichiren, who founded Nichiren Buddhism.

Christianity was introduced into Japan first by Jesuit and then by Franciscan missionaries in the mid- to late 16th century. After the establishment of the Tokugawa shogunate (1603), Christians were persecuted, and Christianity was totally banned in the 1630s. Inaccessible and isolated islands and the peninsula of western Kyushu continued to harbour "hiding Christian" villages until the ban was lifted by the Meiji government in 1873. Christianity was reintroduced by Western missionaries, who established a number of Russian Orthodox, Roman Catholic, and Protestant congregations. Practicing Christians account for only a tiny fraction of the total population.

The great majority of what are now called the "new religions" were founded after the mid-19th century. Most have their roots in Shintō and shamanism, but they also were influenced by Buddhism, Neo-Confucianism, and Christianity. One of the largest, the Sōka Gakkai ("Value Creation Society"), is based on a sect of Nichiren Buddhism. Another new Nichiren sect to attract a large following is the Risshō Kōsei-kai. New Shintō cults include Tenrikyō and Konkōkyō.

DEMOGRAPHIC TRENDS

Japan's population distribution is highly variable. The mountainous character of the country has caused the population to concentrate within the limited plains and lowlands—notably along the Pacific littoral. The increased population, however, has been absorbed into the ever-expanding urban areas, while the population of rural districts has declined considerably; this has had the effect of further concentrating population in a limited area.

Population growth trends

Japan experienced spectacular population growth after 1868, the population increasing nearly fourfold since then. This increase is directly related to slow but steady urban growth; the development of Hokkaido, Tōhoku, and southern Kyushu; and the introduction of commercial agriculture. In 1897, when industrialization first began, the population numbered more than 42 million. From 1898 to 1918 growing industrial cities and mining towns absorbed a large population, as did Hokkaido and the sericultural (silkworm-raising) rural districts.

In 1920, when the first precise census was conducted, the population was nearly 57 million. Between 1919 and 1945 Tokyo-Yokohama, Ōsaka-Kōbe, Nagoya, and northern Kyushu developed as the nation's four major industrial districts. At the same time, some of the smaller cities lost their ability to sustain a growing population, and some of them declined. By 1940 the population had grown to more than double that of 1868. During World War II there was a marked migration to the rural areas to avoid aerial bombing, and some cities such as Ōsaka were reduced to one-third their previous size. After 1945 the repatriated population of nearly 9 million and the temporarily explosive increase in the birth rate caused abnormally high growth.

The rapid rehabilitation of industry after 1950 resulted in the continued concentration of population in the Pacific coastal areas. Rural areas outside the direct influence of urbanization were subjected to a marked decline. Adult males migrated to the Pacific coast, and many of those who remained at home periodically left as temporary labourers, creating a constant outflow of population from the mountainous areas and isolated islands. In many places, emigration was so marked that the remaining population could not maintain a balanced community, and whole settlements were abandoned. These trends have continued in the 1990s, although rural-to-urban migration has slowed somewhat, and people have been leaving city centres for outlying districts and suburbs.

The striking demographic feature in post-World War II Japan is the decline of birth and death rates, the result of families having fewer children and of health conditions improving markedly. Japan's rate of population increase is now one of the world's lowest and its life expectancy among the highest. Consequently, the country has a rapidly aging population, a circumstance that is creating severe labour shortages for its vast economy. Low-skilled job needs, at least, are being met by a growing number of temporary foreign workers. (Ak.W./Y.M./G.La.)

The economy

GENERAL CONSIDERATIONS

Japan is remarkable for its extraordinarily rapid rate of economic growth in the 20th century, especially after World War II. This growth has been based on unprecedented expansion of industrial production and on an aggressive export trade policy. The emphasis on trade stems from Japan's lack of the natural resources needed to support its industrial economy, notably fossil fuels and most minerals; in addition the limited amount of arable land in the country forces it to import much of its food needs.

Global economic status

Japan is now the world's second largest economic power, ranking only behind the United States. It has developed a highly diversified manufacturing and service economy and is one of the world's largest producers of motor vehicles, steel, and high-technology manufactured goods. Japan's standard of living did not increase as rapidly as did the overall economy in the early postwar decades—in large part because of the high percentage of capital reinvestment in those years—but by the mid-1980s it had caught up and was comparable with that found in other developed countries.

Although Japan now has one of the world's highest per capita gross national products, a marked disparity remains between personal income levels and the development of its housing and transportation infrastructure. This disparity is being addressed somewhat by increased public-sector infrastructure investments.

The role of government. Japan's system of economic management is probably without parallel in the world. The extent of direct state participation in economic activities is limited, and the trend is for even less direct involvement. Nonetheless, the government's control and influence over business is stronger and more pervasive than in most other free-enterprise countries. This control is not exercised through legislation or administrative action but through constant—and to an outsider almost obsessive—consultation with business and through the authorities' deep indirect involvement in banking. Consultation is mainly by means of joint committees and groups that keep under review, monitor the performance of, and set targets for nearly every branch and sector of the economy. In addition there are several agencies and government departments that concern themselves with such aspects of the economy as exports, imports, investment, and prices, as well as with overall economic growth. The most important of these agencies is the Economic Planning Agency, which forms part of the Prime Minister's Office and, apart from monitoring the daily running of the economy, also is responsible for long-term planning.

The system works well, avoiding major crises in government-business relations, because of the unusual self-discipline of Japanese businessmen in their dealings with the authorities and the government's deep understanding of the role, needs, and problems of business. The practice of long-term economic planning is a major force in the functioning of the Japanese economy. The need for large-scale government participation in economic activities is thereby obviated, and, unlike many governments in countries practicing free enterprise, the state appears to be reluctant to extend its direct role. In the 1980s the government relinquished to the private sector its monopolies over the tobacco and salt industries and domestic telephone and telegraph services, and the publicly owned Japanese National Railways was privatized. The government retains an interest in international telecommunications services and radio and television broadcasting. It also remains active in matters deemed to be of strategic interest, notably nuclear-power generation, which is subsidized through a major program to increase generating capacity.

The government's role in banking

The government's economic influence is supplemented by its substantial role in banking. The state owns a number of financial institutions, such as the Japan Development Bank, the Export-Import Bank, the Small Business Finance Corporation, and the Housing Loan Corporation, whose principal objectives are to finance private enterprise in areas that are considered particularly desirable. The Ministry of Finance and the Bank of Japan have considerable influence over business investment decisions because of the close interdependence of business, the commercial banks, and the central bank.

Taxation. Tax revenues account for a great majority of the government's total income. Taxes on individuals and on business constitute most of the tax revenues. The largest source from individuals is the progressive income tax. The tax burden on corporations is high, exceeding 30 percent for larger firms. In the late 1980s the tax structure was augmented with a consumption (value-added) tax on most goods and services.

Trade unions and employers' associations. Japanese trade unions have had a relatively short history. Although there were several labour organizations before World War II, trade unions became important only after the U.S. occupation forces introduced legislation that gave workers the right to organize, to bargain with employers, and to strike. Because Japanese trade unions were generally organized on a plant or enterprise basis, their number was relatively large, and in many cases there were different organizations for different plants of the same company.

The great majority of the enterprise unions became affiliated to federations that were loosely organized on craft lines. Most of these, in turn, became affiliated with one of the four major national labour organizations established after the war: the left-wing and highly political General Council of Trade Unions of Japan (Sōhyō), the more moderate and less political Japan Confederation of Labour (Dōmei), the National Federation of Industrial Organizations (Shinsambetsu), and the Federation of Independent Unions (Chūritsu Rōren). Sōhyō was the largest of the four, and Dōmei was its principal rival; Chūritsu Rōren often associated itself with Sōhyō, especially during the annual "spring offensive" (*shuntō*) wage drive.

Interest in uniting the rival national organizations deepened during the 1980s mainly because of the trend toward ever greater concentration in industry and greater cooperation among the various employers' organizations. There also was a growing feeling that Sōhyō's emphasis on ideology was no longer adequate. In the late 1980s the major national organizations and other private- and public-sector unions were reorganized into the Japanese Trade Union Confederation (Rengō); those unions politically more to the left of Rengō formed the much smaller National Confederation of Trade Unions (Zenrōren).

Rengō

The national labour organizations are concerned mainly with such questions as wages, prices, and working conditions. Individual enterprise unions, however, retain much of their independence in dealing with employers. While the craft and national federations formulate general policy, discuss and advise on strategy, and coordinate wage offensives, serious negotiations are usually conducted by individual unions and the employees. One result of Japan's industrial, as opposed to craft, unionism is that demarcation disputes and interunion rivalry for members are relatively rare. Furthermore, if judged in terms of working days lost, Japanese labour relations have been noticeably more amicable than those in other developed countries, such as the United Kingdom, Italy, and the United States.

Japan has a well-developed system of chambers of commerce and trade and industry associations. Among the best-known are the Japan External Trade Organization (JETRO) and the Japan Federation of Economic Organizations (Keidanren).

RESOURCES

Minerals. With few exceptions, Japan's mineral reserves are small, especially in fossil fuels; and although its mineral deposits are fairly diverse, the quality of those mined is often poor. Coal, iron ore, zinc, lead, copper, sulfur, gold and silver are among the most important minerals, with lesser quantities of tungsten, chromite, and manganese. Japan also has large deposits of limestone. There is an almost complete lack of nickel, cobalt, bauxite (the ore of aluminum), nitrates, rock salt, potash, phosphates, and crude petroleum and natural gas.

Coal is the country's most important mineral, although it is of relatively poor quality and is difficult to mine. Reserves are concentrated in Hokkaido and Kyushu. Oil deposits are meagre, domestic oil production accounting for a negligible fraction of Japan's oil consumption. The main oil- and gas-bearing belt extends from northern Honshu on the Sea of Japan to the Ishikari-Yūfutsu lowlands in Hokkaido. Natural-gas reserves also have been found in eastern Chiba prefecture and offshore east of Tōhoku.

Although Japan ranks as one of the world's major steelmakers, domestic resources and production of iron ore are small; Japan imports almost all of its iron ore. Japanese ore is of poor quality and is obtained mostly from small mines in northern and western Honshu. Reserves of copper, once Japan's most important metallic ore, are nearly depleted, and the bulk of copper ore is imported. Lead and zinc are often found in conjunction with copper.

Biological resources. Because of the country's mountainous terrain, the supply of agricultural land is limited. Japan's largely infertile and immature soils require careful husbandry and fertilization. Timber resources are extensive, consisting of broad-leaved and coniferous forests, but much of the forestland is located in inaccessible mountain areas. Most of the forest area is privately owned, and much of it is distributed among a large number of relatively small holders. The rest is publicly owned; large-scale reforestation has taken place in these areas.

Forests

Water resources. Japan's relatively wet climate provides the country with considerable freshwater supplies. The general reliability of the precipitation pattern, coupled with Japan's extensive network of rivers that can be used for irrigation, make possible extensive wet-rice cultivation; flooding, however, is a serious problem in many parts of the country. As a result of the mountainous terrain the country's ample hydroelectric potential is distributed in an uneven fashion. In addition, most hydroelectric power plants cannot operate at full capacity for more than a few months of the year because of seasonal variations of rainfall and the difficulty of constructing adequate storage facilities.

AGRICULTURE, FORESTRY, AND FISHING

Agriculture. Agricultural production is stable or actually declining and accounts for a small and decreasing proportion of the national income. Despite rapid increases in yields after World War II, agricultural productivity per person now is considerably less than per capita output in other sectors of the economy. The agricultural sector employs a relatively large proportion of the working population in comparison to its contribution to national income, but the percentage of the workforce in agriculture is dropping. Many farmers are leaving agriculture for manufacturing and service-industry employment, and most others have to rely on outside occupations for a substantial part of their income.

Japanese agriculture is characterized by a large number of small and often inefficient farms. Larger farms generally are found in Hokkaido, where units of 25 to 50 acres (10 to 20 hectares) are not uncommon. The country's principal crop is rice. Other important farm products include wheat, barley, potatoes, fruits, vegetables, and tea.

The government's agricultural policy has encouraged self-sufficiency in the more important commodities (although

Farmworkers receiving instruction on how to pick and package lettuce, Iwate prefecture, Honshu.
© R. Ian Lloyd/The Hutchison Library, London

this has been achieved only for rice), enlarging the size of the average holding, and closing the gap between rural and urban incomes. The central feature of this policy has been an artificially high producer price for rice. This has succeeded in raising farm incomes and has led to increases in rice production. A surplus of rice, eventual domestic consumer resistance to high prices for rice and other supported commodities, and pressure from foreign governments to reduce barriers to agricultral trade with Japan threaten to undermine this policy.

Livestock-raising, an important farming activity, is generally practiced on a small scale; the largest dairy and beef-cattle herds are in Hokkaido. Most feeds must be imported, and production costs are high.

Livestock

Forestry and fishing. Despite Japan's considerable forest cover, forestry is a marginal activity. In part this is because many of the best stands are inaccessible. In addition, the reforestation in areas that were excessively logged during World War II is not yet available, and Japan's huge demand for timber, high labour costs and other inefficiencies, and concerns about erosion in Japan's steeply sloped mountains plague the domestic logging industry.

Japan relies heavily on the sea as a source of food, and it has one of the largest fish catches of any nation in the world. Much of the catch is derived from long-distance deep-sea fisheries. In spite of its dominant international position, the Japanese fishing industry faces some serious problems. Local fisheries are depleted by overfishing and water pollution, especially in the Inland Sea, while deep-sea fishing must contend with restrictions placed upon it by nations that have claimed a 200-mile economic zone in their coastal waters. Imports of fishery products exceed exports, the result of the growing Japanese demand for seafood and the high value of the Japanese currency. Aquaculture of fish, shellfish (notably clams and oysters), and seaweed is of increasing importance.

INDUSTRY

Mining and quarrying. Mining is an unimportant and declining branch of the economy. The extractive industry is characterized by a large number of small and relatively inefficient mines that do not lend themselves to the application of modern, large-scale mining methods. Coal, the most important mineral mined throughout most of Japan's industrial period, is now extracted as a marginal operation. Most of the remaining production is in Hokkaido. Virtually the whole of the country's output of oil and natural gas comes from Niigata prefecture. Natural gas also is produced in Chiba and Fukushima prefectures. Zinc and lead production is concentrated in southwestern Hokkaido, northern Tōhoku, and Hokuriku, with the first two areas also associated with sulfur mining. Other metallic ores of economic significance include gold, silver, and tungsten. Limestone quarrying is widespread throughout the Japanese archipelago.

Power. Despite Japan's rapidly increasing consumption of energy, per capita consumption remains considerably lower than that in other industrialized countries. The largest single source of energy is oil; almost the entire demand is satisfied through imports, an important share of which comes from fields developed by Japanese companies. Coal, largely imported, is steadily decreasing in importance. The growth in gas production is greatest for natural gas and liquefied natural gas, which account for the largest share of total production.

Most of Japan's total electric power is generated by thermal plants. Oil is most important, but generation by coal-fired plants also is significant. Nuclear-power generation has been heavily developed since the 1970s, with a number of nuclear plants now in operation. Hydroelectric development is largely concentrated in central Honshu, in Tōhoku, and in some parts of Kyushu. This pattern of distribution ensures that Japan's hydroelectric capabilities are well located in relation to the important industrial areas. Although there is significant undeveloped potential, the best sites are already utilized, and further additions to capacity are increasingly expensive.

Hydroelectricity

Manufacturing. The most notable feature of Japan's economic growth since World War II is the rapid development of manufacturing, with progress in quantitative growth, quality, variety, and efficiency. Japan is a sometimes feared competitor whose products are in great demand worldwide. It is one of the world's principal shipbuilders and automakers and is a major producer of such basic products as crude steel, synthetic rubber, aluminum, sulfuric acid, plastics, cement, pulp and paper, refined copper, and cotton yarn. It has some of the world's largest and most advanced industrial plants. The most spectacular growth has been in the production of motor vehicles, iron and steel, machinery (including robots), petrochemicals, precision equipment (notably cameras), and advanced electronic products. Some of the older industries, however, have advanced relatively slowly. The lumber and wood industry, textiles, and foodstuffs have failed to match the expansion in manufacturing as a whole.

A principal reason for Japan's postwar industrial performance was the high level and rapid growth of capital investment, especially in the 1960s and '70s. A boom in equipment investment provided the iron-and-steel and machine-building industries with a rapidly growing home market, allowed for a spectacular increase in productive capacity and in the scale of operations, and led to a rapid replacement of old machinery. This, in turn, resulted in a considerable improvement in productivity throughout the economy and enabled industry to grow, despite an acute shortage of skilled labour and rising wages. Thus, a high rate of labour productivity relative to other major industrial countries, high-quality products, and the extensive use of technological innovations gave many sectors of Japanese manufacturing a formidable advantage over their rivals, with the result that the country's exports soared. A subsequent strategy, carried out with particular success by manufacturers of automobiles and advanced electronic products, was to set up overseas facilities in parts of Asia, North America, and Europe.

The *keiretsu* system

Japanese industry is increasingly characterized by a tendency toward tie-ups, mergers, and takeovers among the larger manufacturing and industrial concerns. The much-studied and controversial *keiretsu* system illustrates this feature of the modern Japenese economy. These groups of companies provide a competitive edge to Japanese firms by managing the risks of manufacturing, distribution, and sales through extensive crossholding of company stocks, long-range strategies aimed at garnering market share without regard to short-term profit, and control over retail prices. These actions are made possible by the gradual relaxation and the increasingly flexible interpretation of the country's antimonopoly laws enacted after World War II. The authorities accept the argument that greater concentration at the top is essential in order to improve efficiency, to make better use of the existing resources, and to increase or maintain international competitiveness. The growing competitiveness of foreign products and increases in direct foreign investment both reinforce and challenge this economic perspective.

FINANCE

Japan's complex financial system is different from that of other developed countries in a number of important respects, the two most important being the major role played by banking and the relatively minor position of securities. The Japanese financial establishment became a major international force in the 1980s: Japan's banks came to dominate international banking, while the Tokyo Stock Exchange emerged as one of the largest securities markets in the world. Much of this growth was based on speculation in a "bubble" economy of highly inflated real estate values. The bursting of the bubble in the early 1990s seriously affected both banking and the securities market.

Banking. The Bank of Japan, established in 1882, is the sole bank of issue; it also plays an important role in determining and enforcing the government's economic and financial policies. The bulk of domestic banking business is transacted through commercial banks, of which the city banks are the most important. There are also a number of long-term credit banks, some government financial institutions—including the Japan Development Bank and the Export-Import Bank—and many mutual savings and loan banks and credit associations.

One of the more unusual features of the Japanese financial system is the high degree of interdependence between the central bank, the commercial banks, and industry. Traditionally, industry has relied on banks for a large part of its borrowing requirements, and, although the importance of its own capital has increased, private and government financial institutions still account for a substantial part of the total. Since the commercial banks are responsible for most credit extended to industry, their influence on their client companies is considerable. Their active lending policy also means that their liquidity ratios tend to be low by Western standards and that they are forced to rely on call money (money that is readily available to banks as loans) and on large-scale borrowing from the Bank of Japan. The central bank is thereby in a strong position to influence bank operations and to bring about a quick adjustment in the volume of credit through credit ceilings, moral pressure, and other methods. Other sources of finance that are less susceptible to central bank influence include mutual savings and loan banks, credit associations, life insurance companies, and other nonbank financial institutions.

Securities. The bond market is relatively undeveloped because the government's low, long-term interest rate policy has made bonds unattractive as against the comparatively high level of short-term rates. Individuals and institutional investors tend to buy discount debentures only. Bond buying, therefore, is confined chiefly to banks and other financial institutions, which are expected to purchase government and government-guaranteed bonds according to an unofficial allocation quota. The secondary bond market has been in operation since the mid-1960s, and, although over-the-counter transactions have risen rapidly, a significant proportion of the business consists of trading in financial debentures. A step toward improving the efficiency of the bond market was made in 1993, when the market was partially deregulated; this allowed banks to participate in the corporate market through subsidiaries.

Stock exchanges

There are several stock exchanges in Japan; the two most important, Tokyo and Ōsaka, account for almost all the business. The Tokyo exchange, along with the exchanges in New York and London, has become one of the pillars of a global, 24-hour securities market.

TRADE

External trade. *Exports.* Another outstanding feature of Japan's postwar economic development is the rapid advance in overseas sales, even though the share of exports in the country's gross national product generally remains constant. From the point of view of individual industries and as a generator of growth, however, exports are much more important than their contribution to the national income suggests.

Reasons for this spectacular export performance are the wide variety of Japan's industrial output, the shift to products with a relatively high value added, more advanced sales-promotion techniques, and the country's export competitiveness. For many years the rise in labour productivity offset the rapid rise in wage costs; Japan's export prices thus tended to be lower than those of its principal competitors. Japanese exports face a number of challenges, however, including increasingly strong competition from such developing industrial nations as South Korea and Taiwan, growing protectionist sentiments among Japan's chief trading partners, and the effects of a sharply devalued U.S. dollar relative to the Japanese yen.

Exports of machinery (including electronic equipment) and transport equipment account for the largest component of exports. Other important exports include scientific and optical equipment, metal and metal manufactures, and chemicals. The United States is Japan's largest single customer; East and Southeast Asia, western Europe, and the Middle East are other important export destinations.

Imports. After World War II, Japan established relatively high tariffs and instituted restrictive nontariff barriers for many products in order to protect domestic markets. Consistently high trade surpluses led to mounting pressure by Japan's trading partners—notably the United States—for Japan to open its domestic market to foreign goods. Imports have grown steadily, as Japan's trade structure has become more open. The high value of the yen, however, has not resulted in a dramatic increase in imports.

Because of Japan's meagre natural resources, the bulk of its imports are raw materials, foodstuffs, and fuels. The major components of imported manufactured goods are machinery and allied products and chemicals. Japan's largest suppliers include East and Southeast Asia, the United States, the Middle East, western Europe, and Australia.

Internal trade. Japan has a long-established and complex system of wholesale distribution and retail marketing, characterized by numerous intermediary levels in the distribution of goods and small, often family-run retail outlets. This system, for years threatened by Japan's large department stores, is now being challenged by the growth of supermarket and discount-store chains and by mail-order sales. Sales traditionally are transacted in cash, but charge accounts and credit cards are becoming increasingly popular.

Growth of credit

TRANSPORTATION AND COMMUNICATIONS

Until the latter part of the 19th century, the majority of Japanese people traveled on foot. Vehicular traffic was limited to small wagons, carts, or palanquins (*kago*) carried by men or animals. The first railway was built between Tokyo and Yokohama in 1872, and others soon followed, though the rugged terrain required the construction of many tunnels and bridges. Iron ships were built at about the same time, and modern ports were constructed. Road construction, however, tended to lag behind the development of other means of transport, resulting in the present congestion of most urban areas.

Japan now has one of the world's most developed transport and communications networks. Tokyo, especially, is an incomparably large focus for transportation; also important are the Ōsaka metropolitan area—which includes the three cities of Ōsaka, Kōbe, and Kyōto—and Nagoya. Other cities—notably Kita-Kyūshū, Fukuoka, Sapporo, Sendai, and Hiroshima—function as regional hubs. The largest volume of intercity or interregional transport, in both passengers and goods, moves between the two largest metropolitan regions. Kyushu is connected with Honshu by the world's first undersea railway tunnel (built in 1941), by an undersea double-decked road tunnel (built in 1958), and by a huge suspension bridge (opened in 1973). With the opening in 1988 of a railway tunnel between Hokkaido and Honshu and of a multiple-span railway-road bridge between Honshu and Shikoku, all four of Japan's main islands are now linked by surface transport.

Roads. The development of Japan's road network lags behind the country's general economic progress and is especially inadequate for the large number of cars. Road construction is hampered, however, by the limited area of land in proportion to population. Limited-access highways have been built between major cities and to some scenic areas, and these are being expanded into a nationwide network. The metropolitan regions of Tokyo and Ōsaka have fairly extensive expressway networks within their respective built-up areas. Surface-street patterns in Japanese cities are manifold, however, and often hamper the flow of traffic. Cities such as Kyōto and Nara still preserve the gridiron street pattern of the ancient Chinese city plan, though with modifications in built-up inner parts of the cities. In many rural areas as well, the ancient pattern of land division and the resultant road pattern take the rectangular gridiron form. Feudal towns, especially fortified (castle) towns, may have somewhat similar street patterns, though in many cases these are modified (generally in the form of concentric rings) to follow former defensive lines.

Japan has an extremely high density of motor vehicles per unit area in the plains and in other inhabited areas. Trucks represent a much higher proportion of vehicular traffic than in other major motorized countries. The great bulk of domestic freight transport is by truck. Many families now have two or more automobiles and are more likely to drive to a destination than in the past, resulting in road congestion in the big cities and in industrial areas. Although railways still play the major role in carrying

Road congestion

Orion Press, Tokyo

The multiple-span Seto Great Bridge over the Inland Sea, linking Kojima, Honshu, with Sakaide, Shikoku.

commuters, there appears to be no practical solution to the problem of how to reduce the number of cars on the roads. The increases in noxious exhaust gases and in the noise of the traffic are serious problems. Steps taken to alleviate them include stringent pollution-control standards for automobiles and the installation of noise barriers on highways in densely populated areas.

Railways. Railways play an extremely important role in passenger travel, though they continue to give way to competition from road and air transport. The development of a modern rail network was an early and farsighted goal of the government after the Meiji Restoration (1868). The first streetcar line was constructed in Kyōto in 1891 and used the electricity from the nation's first power station. In subsequent years Japan, unlike most other Asian countries, developed extensive intraurban and suburban railroad systems; the period between the two world wars, in particular, saw the construction of many railroad lines to the suburbs to serve the needs of growing numbers of middle-income people. In 1927 the first subway was built in Tokyo's downtown district, and over time it was expanded into one of the most extensive systems in the world. Subways subsequently were built in most of Japan's largest cities.

The jewel of Japan's extensive passenger-rail system is the high-speed Shinkansen. Originally a part of the government-owned Japanese National Railways and now one of the companies of the Japan Railways (JR) Group, it began operations in 1964 on the New Tōkaidō Line. Named for the Tōkaidō, the ancient highway between Kyōto and Tokyo, the line provides frequent service on an electrified, double-track route between Tokyo and Ōsaka. It is part of what eventually will be a nationwide network of high-speed trains linking all major cities. Following completion of the New Tōkaidō Line, Shinkansen service was extended westward to Okayama in 1972 and then to Fukuoka on Kyushu in 1975. Two lines radiating outward from Tokyo—north to Niigata and northeast to Morioka—were opened in 1982.

The most serious traffic problem is caused by congestion on commuter-rail transport within the large cities. Most commuter trains are very crowded during rush hours, with some trains carrying many more than the number of passengers for which they were designed. Services are being expanded to cope with the growing demand.

Port facilities. Japan is one of the world's principal seagoing nations and has one of the world's largest merchant fleets. It has engaged in seafaring since early times; but large modern trading ports were not developed until the second half of the 19th century, after the country had reopened to foreign trade following a period of near isolation from the rest of the world. The first of these, Yokohama and Kōbe, remain Japan's leading trade entrepôts, the former being the outport of Tokyo and the latter the outport for Ōsaka and Kyōto. Other important modern ports include Chiba, Nagoya, Kawasaki, Kita-Kyūshū, Mizushima, and Sakai.

Air transport. Both domestic and international air transportation are important in Japan. Before World War II, air transportation was considerably restricted, but, since the foundation in 1953 of Japan Air Lines (JAL), international flights have increased manyfold. Despite competition by railways, especially the Shinkansen, the volume of domestic air transport continues to increase. Tokyo is the main centre of the nation's domestic and international air travel, followed by Ōsaka. Other major airports are in Nagoya, Sapporo, and Fukuoka. All other metropolitan areas in Japan are connected by air routes. Generally speaking, southwestern Japan is covered by a denser network of air transport than other regions, primarily because of the presence of many islands.

Major airports

Communications. The Japanese network of telecommunications and of postal services is among the best in the world. The hundreds of islands, as well as the remotest villages deep in the mountains, are effectively linked by these services. Japan is now a world leader in the use of advanced telecommunications, including widespread use of facsimile transmissions and electronic-mail systems. Per capita telephone ownership is high. The government's privatization of portions of the telecommunications industry—creating Nippon Telegraph and Telephone (NTT), one of the largest private firms in the world—is noteworthy. NTT, in concert with the national government, is creating one of the world's most sophisticated domestic and international communications systems, including a national telecommunications network that incorporates fibre-optic transmission capabilities.

Administration and social conditions

GOVERNMENT

Constitutional framework. Japan's constitution was promulgated in 1946 and came into force in 1947, superseding the Meiji Constitution of 1889. It differs from the earlier document in the following points: the emperor, rather than being the embodiment of all sovereign authority (as he was previously), is the symbol of the state and of the unity of the people, while sovereign power rests with the people; Japan renounces war as a sovereign right; and fundamental human rights are explicitly guaranteed. Furthermore, the government is now based on a constitution that aims at maintaining Japan as a peaceful and democratic country in perpetuity.

The role of the emperor

The emperor's major role now consists of such formalities as appointing the prime minister—who is first designated by the Diet (Kokkai)—and appointing the chief justice of the Supreme Court (Saikō Saibansho), convoking sessions of the Diet, promulgating laws and treaties, and awarding state honours—all with the advice and approval of the Cabinet (Naikaku).

Legislative powers are vested in the Diet, which is popularly elected and consists of two houses. The House of Representatives, or lower house (Shūgiin), ultimately takes precedence over the House of Councillors, or upper house (Sangiin). Membership in the House of Representatives is based on proportional representation from prefectural districts, while that in the House of Councillors is divided between proportional representation and at-large representation. The House of Representatives controls the budget and approves treaties with foreign powers. Executive power is vested in the Cabinet, which is organized and headed by the prime minister. If the House of Representatives passes a resolution of no-confidence or refuses to pass a vote of confidence in the government, the Cabinet must resign, unless the House of Representatives is dissolved within 10 days. There are governmental ministries and agencies in addition to the Prime Minister's Office. All offices of the central government are located in and around the Kasumigaseki district in central Tokyo. An independent constitutional body called the Board of Audit is responsible for the annual auditing of the accounts of the state.

Local government. Japan is divided into 47 prefectures, 43 of which are *ken* (prefectures proper), one of which (Tokyo) is a *to* (metropolitan prefecture); one (Hokkaido) is a *dō* (district), and two (Ōsaka and Kyōto) are *fu* (urban prefectures). Prefectures, which are administered by governors and assemblies, vary considerably both in area and in population. A prefecture is further subdivided into minor civil divisions; these include the city (*shi*), town (*machi* or *chō*), and village (*mura* or *son*). All these local government units have their own mayors, or chiefs, and assemblies. Before World War II there were also counties (*gun*), consisting of towns and villages but excluding cities within a prefecture. This county system survives only in the form of statistical units.

An intermediate level of governmental services is formed between the central and prefecture levels. The branch offices of several central ministries are located in certain cities, which—as regional centres—generally administer several prefectures together. Designated cities (*shitei toshi*), which must have populations of at least 500,000 each, are divided into wards (*ku*). A ward has a chief and an assembly, the former being nominated by the mayor and the latter elected by the residents. Tokyo has 23 special wards (*tokubetsu ku*), the chiefs of which are elected by the residents. These special wards, created after the metropolitan prefecture was established in 1943, demarcate the city of Tokyo from the other cities and towns that make up the metropolitan prefecture; the city proper, however, no longer exists as an administrative unit.

The political process. *Elections.* Japan has universal adult suffrage for all citizens 20 years of age or older. Members of both the House of Representatives and the House of Councillors are chosen by general elections. The 511 members of the House of Representatives are elected to four-year terms, but this may be terminated earlier if the house is dissolved. The 252 members of the House of Councillors are elected to six-year terms, with half the members being elected every three years. Members of the House of Representatives are elected from 130 multiple-member constituencies. The electoral procedure for the House of Councillors differs from that for the lower house in that about two-fifths of the total are elected from a national constituency, in which each voter casts a vote for a national candidate; the remaining members are elected from the prefectural constituencies. Heads of local governmental units, such as prefectures, cities, special wards, towns, and villages, are elected by local residents.

The number of seats for each Diet constituency was determined largely on the basis of the population in each area in 1947, with some modifications resulting from the population increase in urban constituencies. Nonetheless, by the early 1980s Japan's population distribution had changed so much that the value of a vote in a sparsely populated rural district might be five times that of one in an urban district. A limited amount of reapportionment was done in the mid-1980s, which redressed somewhat this imbalance, and in 1994 legislation was passed that would reduce the size of the lower house to 500 and create 300 single-member constituencies, the remaining members being elected from electoral districts based on proportional representation.

Electoral reform

Political parties. Party politics in Japan was inaugurated during the Meiji period (1868–1912) and subsequently suppressed during the war years of the 1930s and '40s. The freedom to organize political parties was subsequently guaranteed by the 1947 constitution. Any organization that supported a candidate for political office was required to be registered as a political party; thousands of parties, most of them of local or regional significance, have since been organized, merged, or dissolved.

Several parties rose to national prominence. Chief among these was the Liberal-Democratic Party (LDP), generally conservative and pro-business and the dominant force in government for nearly all of the second half of the 20th century. The Social Democratic Party of Japan (SDPJ), until 1991 called the Japan Socialist Party (JSP), long was the major opposition party, drawing much of its support from labour unions and inhabitants of the large cities. The moderately socialist Clean Government Party (Kōmeitō), also an important opposition party, drew its main support from the Sōka Gakkai, although the religious organization subsequently renounced any formal ties with the party. The Democratic Socialist Party (DSP) was formed in 1960 by a right-wing splinter group of the JSP. Finally, the Japan Communist Party (JCP), small but influential for its size, remained on the fringe of the opposition.

The Japanese political landscape underwent radical change after 1990. The LDP's long-held control of the government was broken in 1993, the result of a series of corruption scandals in the party and the subsequent resignations of a large number of LDP members to form new parties. The demise of LDP hegemony produced political uncertainty, notably the unprecedented spectacle of four different prime ministers within a 12-month period in 1993–94. In addition to the formation of new political parties, considerable party realignment occurred, producing an unlikely SDPJ-LDP governing coalition and a large, multiparty opposition coalition called the New Frontier Party (Shinshintō).

Justice. The judiciary is completely independent of the executive and legislative branches of the government. The judiciary system consists of the Supreme Court, eight high courts, a district court in each prefecture (with the exception of Hokkaido, which has four) and many summary (informal) courts. Family courts also are numerous.

The Supreme Court consists of one chief justice and 14 other justices. The chief justice is appointed by the emperor upon designation by the Cabinet, while the other justices are appointed by the Cabinet. The appointment of the justices of the Supreme Court is subject to review in a national referendum, first at the time of the general election following their appointment and then at the general election every 10 years thereafter. An impeachment system also exists; the court of impeachment consists of members of the House of Representatives and of the House of Councillors. The Supreme Court determines questions of the constitutionality of any law, order, regulation, or official act. Lower-court judges are appointed by the Cabinet from a list of persons nominated by the Supreme Court. The appointment term is for 10 years, and reappointment is allowed. All judges of lower courts must retire at the age of 70, according to law.

Security. *Armed forces.* As mentioned above, Japan's 1947 constitution stipulates that the country cannot maintain armed forces for purposes of aggression. Between 1945 and 1950, Japan had no armed forces except for police. After the outbreak of the Korean War, however, the government, at the suggestion of the Allied occupation forces, established a National Police Reserve, which later became the Self-Defense Forces (SDF; Jieitai). The SDF consist of ground, maritime, and air branches, under the civilian-controlled National Defense Council.

The Self-Defense Forces

National defense also is maintained by the collective security system in which the United States participates. Through the Treaty of Mutual Cooperation and Security, concluded between Japan and the United States in 1960 and reaffirmed in 1970, the United States operates military bases in many parts of Japan. The treaty may be terminated one year after either signatory indicates such an intention.

Police. Japan's police services are under the supervision and control of the National Police Agency. Police services operate relatively smoothly; many problems that plague other countries are absent because of Japan's insularity, its nearly uniform ethnic composition, and the high professional level to which its police are trained. The relatively low ratio of extremely violent crimes to total crimes stands in contrast to that of most economically advanced countries. There also are a number of more specialized policing bodies, the largest of which, the Maritime Safety Agency, patrols Japan's coastal waters.

EDUCATION

Japan's modern education system is one of the key elements in the country's emergence as a highly industrialized nation. Businesses long have recognized the value of broadly trained workers, and individuals have seen education as the means to achieve personal advancement. In the Japanese system, where one attends school largely determines one's ultimate social status and financial success. As a result, students from a young age work extremely hard to qualify for the best possible schools.

Higher education is greatly desired. The two great former imperial universities—Tokyo and Kyōto—represent the pinnacle of academic success, and their graduates are considered the best prospects by public and private employers. The high-school curriculum, largely designed as preparation for the difficult and highly competitive university entrance examinations, is especially rigorous. In addition, most students attend one of the large number of extracurricular "cram" schools (*juku*) that help them prepare for the examinations. High-school graduates who do not pass the examinations on their first attempt often study intensively for a year and retake the tests; a growing number of passed-over candidates, however, pursue undergraduate degrees abroad.

Entrance examinations

Development of the modern system. Many educational institutions existed even in the feudal period preceding the Meiji Restoration of 1868. Numerous private temple schools (*terakoya*), mostly in towns, functioned as elementary schools. Provincial lords (daimyo) also established special schools for children of the warrior class. Yet another type of school instructed primarily the children of wealthier merchants and farmers.

The modern educational system was introduced immediately after the Meiji Restoration. The government set up elementary and secondary schools throughout Japan in 1872, and in 1886 a system providing three to four years of education was inaugurated. The introduction of modern education did not encounter many problems, primarily because it utilized the existing system. Free compulsory education was introduced in 1900, and in 1908 it was extended to a period of six years. Since 1947, education has been compulsory for a nine-year period, beginning at the age of six.

System organization. *Primary and secondary education.* The primary and secondary educational systems are organized as follows: kindergarten (not compulsory), from one to three years; compulsory elementary school, six years; compulsory middle school, three years; and high school (not compulsory), another three years. Public elementary and middle schools are free, and there are numerous private institutions. Japan is one of the few countries in the world that provide a complete and thorough education for almost all their people. Although neither kindergartens nor high schools are compulsory, attendance at both has become virtually universal.

In principle, educational administration is decentralized; responsibilities for the budget, curriculum, teacher appointments, and the supervision of elementary and middle schools are in the hands of local educational boards, with the Ministry of Education playing a coordinating role. In practice, however, the ministry keeps a tight rein on curriculum and other aspects of primary and secondary instruction.

Higher education. Institutions of higher education—of which there are now more than 1,100—consist of junior colleges, lasting for two to three years, and ordinary colleges and universities, lasting for four years. In addition, there are five-year technological colleges that combine high school and junior college education. The Tokyo metropolitan area, including Yokohama and many other satellite cities, has a high concentration of both institutions and college students. Of note is the Tsukuba Science City, located about 40 miles northeast of Tokyo, which consists of government research facilities and two universities.

Continuing education. The strong Japanese sense of self-improvement through education extends beyond formal schooling. For example, although most *juku* are geared toward preparation for entrance examinations, many also provide music, art, and sports instruction for school-age children. Adult continuing education is conducted by both local governments and private institutions, offering classes in general education, vocational training, technology, homemaking, home economics, arts, physical education, and recreation. In 1985 the University of the Air began operation as a means of providing opportunities for higher education via television broadcasts.

University of the Air

SOCIAL CONDITIONS

Health. Japan's high standard of living contributes much to the general good health of the country's population. The nation has one of the most comprehensive health-care systems in the world, characterized by national health insurance, extremely low infant-mortality rates, and one of the world's highest life-expectancy rates. Infectious diseases such as typhoid, diphtheria, and scarlet fever have been virtually eliminated, while tuberculosis and dysentery are much less prevalent than they once were. Instead, there have been increases in the diseases associated with advanced industrial societies: cancer, stroke, high blood pressure, heart ailments, and pulmonary diseases are now principal causes of death, as are traffic accidents and suicide. The government has established hundreds of health centres throughout Japan. Most of the country's hospitals are operated by unions, associations, or individuals, and the remainder by local governments and the national government. Japanese medical practice is usually of the Western type, but classical Chinese techniques are also used.

The Japanese people enjoy an increasingly varied diet. Traditional Japanese foods are being supplemented or replaced partly by Western types of food (notably red meats and dairy products) and partly by Chinese food—to such a degree that the average Japanese no longer regards Western or Chinese food as alien. Although calorie consumption is generally lower than that of Europeans or Americans, overnutrition causing excess weight is considerably more prevalent than in the past.

Welfare. The vast discrepancies that existed between the conditions of the rich and poor before World War II have been reduced, largely as a result of the agricultural land reforms between 1946 and 1950 and of the application of a graduated income tax. The great majority of Japanese now regard themselves as middle class, although within this designation there still are considerable differences in income levels and property ownership. Most of those in the upper middle-income group own their own homes, usually houses with several rooms surrounded by a garden; those in the lower middle-income group usually live in a two- to five-room house or (more commonly in urban areas) in an apartment house.

Social-welfare services have expanded considerably since the end of World War II. Programs include social insurance (health insurance, pension insurance, unemployment insurance, and worker's accident compensation insurance), services for the elderly and the physically and mentally handicapped, and care for disadvantaged children. The health-insurance system, established in 1961, covers all Japanese people. The scale of payments into it varies, and in some cases no payments are required. Elderly people may receive many services, including medical examinations, home-help services, recreational services, and institutional care, as well as varying amounts of financial aid. Local governments are obliged to provide welfare services for the physically and mentally handicapped. Various children's welfare programs also exist—for example, medical-care services are free to expectant mothers and to young children from low-income families.

Health insurance

Housing. Despite a considerable increase in the number of single-family homes and apartments, Japan continues to have serious housing shortages in large metropolitan areas. The primary cause of this is high urban population concentrations, creating land prices and housing costs that far exceed average incomes. Other contributing factors include the great destruction of urban housing stock during World War II; the losses caused by earthquakes, typhoons, and floods triggered by heavy rains, and a resultant pattern of housing designed to be replaced relatively frequently; and an increase in the number of families through the breakup of extended families into smaller units. In addition, Japan's overall affluence fuels a demand for larger and better houses and for a redirection of capital investment from the heavy emphasis on industrial expansion to house construction.

The absence of strict zoning in urban areas contributes to the mixed land uses characteristic of Japan's cities; thus, functionally different establishments, such as shops, factories, or houses, are found adjoining one another. In addition, even the largest city may contain scattered plots of agricultural land. These factors, plus the high cost of housing in city centres, are forcing people farther into outlying areas. As a result, the length of daily commuting is steadily increasing, and, not uncommonly, commuters travel two or more hours each way.

To cope with the initial postwar housing shortage, a semigovernmental agency, the Housing Loan Corporation, was established in 1950 to finance house construction at low interest rates. In 1955 another semigovernmental agency, the Japan Housing Corporation (since 1981 called the Housing and Urban Development Corporation) was organized, which contributed significantly to housing construction. In addition, local governments have built a number of units, mostly of the apartment-house type and primarily for low-income families, and many large corporations maintain low-cost apartment or dormitory-style housing for their employees.

Cultural life

THE CULTURAL MILIEU

Influences. Japanese cultural distinctiveness and the manner in which it developed are instructive in understanding how it is that Japan came to be the first and so far only non-Western country to attain great-power status. The Japanese long have been intensely aware of and have responded with great curiosity to powerful outside influences, first from the Asian mainland (notably China) and more recently from the Western world. Japan has followed a cycle of selectively absorbing foreign cultural values and institutions and then adapting these to existing indigenous patterns, this latter process often occurring during periods of relative political isolation. Thus, outside influences were assimilated, but the basic sense of Japaneseness was unaffected; for example, Buddhist deities were adopted into the Shintō pantheon. Japan's effort to modernize quickly in the late 19th and 20th centuries—undertaken at great national and personal sacrifice—was really an extension of the same processes at work in the country for centuries.

Absorption and adaptation

Prehistoric Japanese culture was exposed to ancient Chinese cultural influences beginning some two millennia ago. One consequence of these influences was the imposition of the gridiron system of land division, which long endured. Chinese writing and many other Chinese developments were introduced in the early centuries AD; the writing system underwent many modifications over the centuries, since it did not fit the Japanese language. Buddhism—which originated in India and underwent modification in Central Asia, China, and Korea before reaching Japan about the 6th century—also exerted a profound influence on Japanese cultural life, although over the course of time it was modified profoundly from its antecedent forms. Similarly, Chinese urban design was introduced in the layouts of the ancient capital cities of Nara and Kyōto but did not proliferate in the archipelago.

The Japanization of introduced cultural elements was greatly accelerated during the 250-year period of near-isolation that ended in the mid-19th century. After the Meiji Restoration, Japan began to modernize and to industrialize on the European and American pattern. In the period since then, the United States generally has exerted a more conspicuous influence on Japanese cultural and social life than has Europe. Western cultural traits have been introduced on a large scale through the schools and the mass communication media. Western scientific and technical terms have been widely diffused in translation and have even been reexported to China and Korea. American and European influences on Japanese culture are in evidence in literature, the visual arts, music, education, science, recreation, and ideology.

Modernization was accompanied by cultural changes. Rationalism and socialism based on Christianity, as well as Marxism, became inseparably related to everyday Japanese life. Western or Westernized music seems to be preferred to traditional Japanese music at most social levels. Although Japanese Christians form a tiny percentage of the population, Christmas (or the outer trappings of it) is widely observed. The use of Western dress among the Japanese, in place of the traditional kimono, is widespread, although women tend to wear formal kimonos at celebrations, and both men and women may use casual styles for home wear. House construction also has been changed considerably by the introduction of Western architectural forms and functions. In shape, in colour, and in building materials, many contemporary Japanese houses are significantly different from the traditional ones; they now have more modernistic shapes, use more colours, and are more often made of concrete and stucco.

Aesthetics. The dual influences of East and West have helped construct a modern Japanese culture that offers familiar elements to the Westerner but that also contains a powerful and distinctive traditional cultural aesthetic. This can be seen, for example, in the intricate detail, miniaturization, and concepts of subtlety that have transformed imported visual art forms. This aesthetic is best captured in the Japanese concept of *shibui* (literally, "astringent"), or refined understatement in all manner of artistic representation. Closely related are the twin ideals of cultivated simplicity and poverty (*wabi*) and of the celebration of that which is old and faded (*sabi*). Underlying all three is the notion of life's transitory and evanescent nature, which is linked to Buddhist thought (particularly Zen) but can be traced to the earliest examples of Japanese literature.

Concepts of *shibui*, *wabi*, and *sabi*

THE ARTS

Delicacy and exquisiteness of form, together with simplicity, characterize traditional Japanese artistic taste. The Japanese tend to view the traditional Chinese arts generally as being too grandiose or showy. The more recently introduced Western arts are felt to suffer from flaws of exuberant self-realization at the expense of earnest exploration of the conflicts in human relations, in particular the notions of divided loyalties between community, family, and self that create the bittersweet melancholy so pervasive in Japanese traditional drama.

Traditional forms. The highly refined traditional arts of Japan include ikebana (flower arranging), the tea ceremony, painting, calligraphy, woodblock printing, dance, music, theatrical plays (including such forms of drama as kabuki, a highly stylized form of drama characterized by singing and dancing; bunraku, the puppet theatre; nō, the classic form of dance-drama), and *gagaku* (court music), gardening, and architecture. Such arts as ikebana, the tea ceremony, and calligraphy are studied and practiced by a great many Japanese. Ikebana and the tea ceremony, in particular, are popular among young unmarried women, since these are regarded as appropriate cultural or aesthetic accomplishments for future housewives. Traditional Japanese painting, dance, and music have, however, lost much of their traditional popularity, though the poetic forms of haiku and waka continue to flourish.

Traditional handicrafts constitute some of Japan's finest examples of visual arts. Notable are the various styles of pottery, laquerwork, cloisonné, and bamboo ware, as well as papermaking, silk weaving, and cloth dyeing.

With the advance of modernization, many folk traditions and forms of folklore are disappearing. The widespread use of standard Japanese has accelerated this trend, since local cultures are directly related to dialects. Folk songs, for example, are generally no longer commonly sung except in some remote areas in northern and southwestern Japan. Folk music and dance are related to local life and are often significantly concerned with the local religion (whether animistic, Shintō, or Buddhist), agriculture, or human relations (including the theme of love). Some, however, still enjoy a great popularity, which has been increased through the mass media.

Western forms. Despite their perceived aesthetic shortcomings, Western art forms have been fully embraced by the Japanese. Major cities often have several symphony or-

Jim Brandenburg—TSW-CLICK/Chicago

Craftsman testing a *shō*, a type of mouth organ made from wood and bamboo that is used in *bugaku* (court music).

chestras, and Western-style painting, sculpture, and architecture are widely practiced. The cinema has been highly successful at taking a Western form and putting it through a Japanese aesthetic filter to produce a distinctive style; internationally acclaimed Japanese film directors include Kurosawa Akira, Ozu Yasujirō, and Mizoguchi Kenji.

CULTURAL INSTITUTIONS

Libraries

The Japanese are among the most literate peoples in the world. The National Diet Library in Tokyo (which also includes branch libraries) is the single largest library in Japan. The concept of the public lending library, however, is fairly new in Japan, which partially explains the country's high incidence of commercial book sales.

Most of Japan's major cultural institutions—including the Japan Academy, the Tokyo National Museum, and the National Theatre—and many of its most prestigious universities—*e.g.*, the public University of Tokyo and private Waseda and Keio universities—are located in Tokyo. Japan's numerous Buddhist temples also contain a great many cultural properties, especially those located in Kyōto and Nara. In addition to the many public institutions, numerous private museums, art galleries, theatres, and gardens are found throughout the country, and Japanese department stores offer free or low-cost exhibitions.

Of particular note is the national government's practice of identifying and recognizing various artists, performers, and artisans of traditional Japanese art forms. Designated "living national treasures," these individuals receive an annual stipend that allows them to practice their skills and to pass them along to apprentices. This program helps preserve many of the forms and styles that otherwise might disappear.

THE MEDIA

The print and broadcast media have long been influential in Japan. Although their activities were somewhat circumscribed by the government until the end of World War II, they now operate in an atmosphere of considerable freedom. Commercial advertising has become an immense industry, and Japan now has the second largest market, after the United States. Television and newspapers are the most important advertising media, with magazine and radio advertising being less significant.

The press. Japan ranks as one of the major book publishing countries in the world, and Tokyo is the centre of the Japanese publishing industry. Several thousand magazines are also published, with more than half of these being weeklies.

The role of newspapers is of great importance. Japan's largest dailies rank among the highest in the world in circulation, and all the large papers are generally considered to maintain high editorial standards. Major newspapers print both morning and evening daily editions, and daily circulation is high; the largest papers each have daily press runs of several million. A number of newspapers have nationwide circulation, and some local papers also have large circulations. Kyodo News Service and Jiji Press are Japan's largest news agencies.

Radio and television. Radio and television are used in Japan far more extensively than in any other Asian country and, indeed, more extensively than in most other countries in the world. Radio broadcasting began in 1925 with the establishment of the Japan Broadcasting Corporation (Nippon Hōsō Kyōkai, or NHK)—a public corporation financed by license fees that are paid by television-set owners. NHK broadcasts quality, commercial-free programming on both radio and television.

NHK

The first television broadcast was made by NHK in 1952. Television stations now broadcast to all parts of Japan, including all the isolated islands. NHK has been broadcasting overseas programs such as "Radio Japan" since 1953; it now produces radio broadcasts in more than 20 languages and provides a limited amount of television broadcasting. Private commercial broadcasting began in 1951 and has gained widespread popularity. There are now several private television networks, many of them owned by newspaper companies. In addition, satellite and cable television reception is common.

POPULAR CULTURE

Daily life. Contemporary Japanese society is decidedly urban. Not only do the vast majority of Japanese live in urban settings, but urban culture is transmitted throughout the country by a mass media concentrated in Tokyo. Young urban Japanese, in particular, increasingly are known for their conspicuous consumption and for their penchant for trends and fads that quickly go in and out of fashion.

Modern (usually Western) popular music has gained a strong foothold in Japan. Jazz, rock, and the blues are enjoyed by the younger generation, along with half-Westernized or half-Japanized folk and popular songs. Many basically Japanese songs are sung to the accompaniment of Western musical instruments; at the same time, many basically Western subjects are treated in Japanese-style drama or song.

Most entertaining is not done at home, in part because of the small size of most Japanese homes. The commercial landscape of most Japanese cities is among the most diverse and service-oriented in the world, where all manner of food, Japanese or otherwise, can be found. Japanese cuisine, which often is served raw or only lightly cooked, is noted for its subtle and delicate flavours.

Social customs. Especially in the more anonymous world of the city, the traditional arranged marriage (*miai-kekkon*) is being replaced by the love match. It is still common for a family friend, relative, or mentor to act as a go-between (*nakodo*), even if the marriage is a love match. The wedding ceremony itself often consists of a curious blend of East and West: a traditional Shintō ceremony, in which the bride and groom wear elaborate kimonos, typically is followed by a Christian-style observance, with the participants in formal Western attire.

Holidays

Japan has 12 national holidays. New Year's Day is traditionally regarded as the most important of these holidays, with millions of people engaging in a kind of pilgrimage to shrines and temples starting at midnight of December 31. For three days thereafter, people visit shrines and temples, their families, and the homes of friends. The occurrence of multiple holidays in late April–early May (popularly called Golden Week) is one of the most popular vacation times for the Japanese, as is the week of the Bon festival in mid-July or mid-August. Many temples and shrines celebrate their own specific festivals, attracting large numbers of people. City, town, and village authorities, as well as local communal bodies, often organize local festivals.

The Japanese have a great fondness for seasonal blossom- and leaf-viewing. Most popular are the cherry blossoms of spring (in many areas, around Golden Week). Each year, the entire country is captivated by the northward progress of the trees—the so-called "cherry blossom front." This is mirrored in the fall, to a lesser degree, by the southward progress of the turning maple leaves.

Recreation. For much of the postwar period Japanese workers did not exploit the full allowance of vacation time allotted to them, but since 1980 the country as a whole has become more leisure-conscious. Japan has an extensive and well-utilized system of national parks, quasi-national parks, and prefectural natural parks. Travel within Japan is widespread. Nonetheless, because of the high price of domestic transportation and lodging, it is increasingly common for Japanese to travel abroad. The cultural capitals of Europe, the American West Coast, nearby South Korea and Hong Kong, as well as Australia and the Pacific islands are favourite destinations.

Popular outdoor recreational activities include hiking, mountaineering, skiing, skating, golf, and swimming, boating, fishing, baseball, tennis, and football (soccer). Professional soccer now rivals baseball for national attention and patronage. Indoor recreations include shogi and go (both similar to chess), mah-jongg, Japanese and Western card games, basketball, volleyball, table tennis, gymnastics, and such martial arts as judo, kendo, karate, and aikido. Sumo wrestling is also practiced, and watched enthusiastically, both indoors and out. (Y.M./G.La.)

For statistical data on the land and people of Japan, see the *Britannica World Data* section in the BRITANNICA BOOK OF THE YEAR.

HISTORY

Ancient Japan to 1185

PREHISTORIC JAPAN

Pre-Ceramic culture. It is not known when humans first settled on the Japanese archipelago. It was long believed that there was no Paleolithic occupation in Japan, but since World War II thousands of sites have been unearthed throughout the country, yielding a wide variety of Paleolithic tools. These include both core tools, made by chipping away the surface of a stone, and flake tools, made by working with a stone flake broken off from a larger piece of stone. There is little doubt that the people who used these implements moved to Japan from the Asian continent. At one stage, land connections via what are now the Korea and Tsushima straits made immigration from the Korean peninsula possible, while another connection, via what are now the Sōya and Tsugaru straits, allowed people to come in from northeastern Asia.

The Paleolithic Period in Japan is variously dated from 30,000 to 10,000 years ago, although the argument has been made for a Lower Paleolithic culture prior to 35,000 BC. Nothing certain is known of the culture of the period, though it seems likely that people lived by hunting and gathering, used fire, and made their homes either in pit-type dwellings or in caves. No bone or horn artifacts of the kind associated with this period in other areas of the world have yet been found in Japan. Since there was no knowledge whatsoever of pottery, the period is referred to as the Pre-Ceramic era.

Climatic changes help to account for the existence of a Mesolithic stage in early Japanese culture, a time when much of the abundant fauna of earlier times became depleted by the expanding human population of the archipelago. The introduction of the bow and arrow is regarded as a local response to a decrease in game available for food.

Jōmon culture (7500 BC to *c.* 250 BC). The Pre-Ceramic era was followed by two better-recorded cultures, the Jōmon and the Yayoi. The former takes its name from a type of pottery found throughout the archipelago; its discoverer, the 19th-century American zoologist Edward S. Morse, called the pottery *jōmon* ("cord marks") to describe the patterns pressed into the clay. A convincing theory dates the period during which Jōmon pottery was used from about 10,000 years ago until the 2nd or 3rd century BC. Of the features common to Neolithic cultures throughout the world—progress from chipped tools to polished tools, the manufacture of pottery, the beginnings of agriculture and pasturage, the development of weaving, and the erection of monuments using massive stones—the first two are prominent features of the Jōmon period, but the remaining three did not appear until the succeeding Yayoi period. Pottery, for example, first appeared in northern Kyushu (the southernmost of the four main Japanese islands) about 10,000 BC, in an era that is sometimes called the "incipient" Jōmon period. Jōmon is thus best described as a Mesolithic culture, while Yayoi is fully Neolithic.

Appearance of pottery

It is customary to regard changes in pottery types as a basis for subdividing the age into six periods: incipient, very early, early, middle, late, and very late. Since Jōmon culture spread over the entire archipelago, it also developed regional differences, and this combination of both chronological and regional variations gives the evolution of Jōmon pottery a high degree of complexity.

Jōmon dwelling sites have been found in various parts of the country. They can be classified into two types: one, the pit-type dwelling, consisted of a shallow pit with a floor of trodden earth and a roof; the other was made by laying a circular or oval floor of clay or stones on the surface of the ground and covering it with a roof. Remains of such dwellings have been found in groups ranging from five or six to several dozen, apparently representing the size of human settlements at the time. Most of these settlements form a horseshoe shape, with a space in the centre that seems to have been used for communal purposes. It can be deduced that each household was made up of several family members and that the settlement made up of such households was led by a headman or shaman.

The people of the Jōmon period lived mainly by hunting and fishing and by gathering edible nuts and roots. The appearance of large settlements from the middle period onward has been interpreted by some scholars as implying the cultivation of certain types of crop—a hypothesis seemingly supported by the fact that the chipped-stone axes of this period are not sharp but seem to have been used for digging soil. Doubtless there was some form of cultivation: starchy yams and taro, probably originating from the continent, were raised, the starch from them formed into a type of bread.

Weaving was still unknown, and archaeological findings indicate that clothes were largely made of bark. Body ornamentation included bracelets made of seashells, earrings of stone or clay, and necklaces and hair ornaments of stone or bone and horn.

No especially elaborate rites of burial evolved, and the dead were buried in a small pit dug near the dwelling. Sometimes the body was buried with its knees drawn up or with a stone clasped to its chest, a procedure that probably had some religious or magical significance. A large number of clay figurines have been found, many representing female forms that were probably magical objects associated with primitive fertility cults.

For years certain scholars have claimed that the bearers of the Jōmon culture were not of Japanese ethnicity but were ancestors of the Ainu, an aboriginal people often regarded as having European (Caucasian) racial connections who now are found in northern Japan. Scientific investigation of the bones of Jōmon people carried out since the beginning of the 20th century, however, has disproved this theory. The Jōmon people might be called proto-Japanese, and they were spread throughout the archipelago. Despite certain variations in character arising from differences in period or place, they seem to have constituted a single ethnic stock with more or less consistent characteristics. The present Japanese people were produced by an admixture of certain strains from the Asian continent and from the South Pacific, together with adaptations made in accordance with environmental changes. Linguistic evidence suggests that a people speaking a language belonging to the Ural-Altaic family moved eastward across Siberia and entered Japan via Sakhalin Island and Hokkaido. Nothing can yet be proved concerning their relationship with the people of the Pre-Ceramic period, but it cannot be asserted that they were entirely unrelated.

The Ainu

The Yayoi period (*c.* 250 BC–*c.* AD 250). The new Yayoi culture that arose in Kyushu, while the Jōmon culture was still undergoing development elsewhere, spread gradually eastward, overwhelming the Jōmon culture as it went, until it reached the northern districts of Honshu (the largest island of Japan). The name Yayoi derives from the name of the district in Tokyo where, in 1884, the unearthing of pottery of this type first drew the attention of scholars. Yayoi pottery was fired at higher temperatures than Jōmon pottery and was turned on wheels. It is distinguished partly by this marked advance in technique and partly by an absence of the proliferating decoration that characterized Jōmon pottery. It developed, in short, as pottery for practical use. It is accompanied by metal objects and is associated with the wet (*i.e.,* irrigated) cultivation of rice. Culturally, the Yayoi represents a notable advance over the Jōmon period.

In China the 3rd and 2nd centuries BC correspond with the period of the unified empire under the Ch'in (221–206 BC) and Han (206 BC–AD 220) dynasties, which already had entered the Iron Age. In 108 BC the armies of the emperor Wu Ti occupied Manchuria and the northern part of the Korean peninsula, where they established Lo-lang and three other colonies. These colonies served as a base for a strong influx of Chinese culture into Korea, whence, in turn, it spread to Japan. The fact that Yayoi culture had iron implements from the outset, and bronze implements somewhat later, probably indicates borrowings from Han culture. Since iron rusts easily, comparatively few objects have been found, but they seem to have been widespread at the time. These include axes, knives, sickles and hoes, arrowheads, and swords. The bronze objects are also varied, including halberds, swords, spears, *taku* (bell-shaped devotional objects from China), and mirrors. The halberds, swords, and spears seem not to have been used in Japan for the practical purposes for which they were evolved in China but rather to have been prized as precious objects.

The cultivation of rice

The cultivation of rice, probably introduced from the Yangtze River delta area of southern China, was one of the most important features of Yayoi culture. The earliest Yayoi pottery and sites, discovered in northern Kyushu, have yielded marks of rice husks as well as carbonized grains of rice; this suggests that rice growing was carried on in Japan from the earliest days of the culture. Traces of paddy fields, their divisions marked with wooden piles, have been found close to sites of settlements in various districts, along with irrigation channels equipped with dams and underdrains, showing that techniques of making and maintaining paddy fields were quite advanced. Rice was first grown in dry (*i.e.*, nonirrigated) fields and marshy areas, however, before paddy cultivation—involving considerable investment of time, labour, and capital—came to predominate.

Generally speaking, the settlements of this period were built on low-lying alluvial land to facilitate the irrigation of the paddies, but at one stage they were built instead in the hills or on high ground. Much as in the Jōmon period, there were two types of dwelling—the pit type and the type built on the surface—but in addition to these, raised-floor structures appeared and were used for storing grain out of the reach of rodents.

With the acquisition of a knowledge of textiles, clothing made great strides compared with the Jōmon period. The cloth was woven on primitive looms using vegetable fibres.

The dead were buried in either large clay urns or heavy stone coffins. Both were common in northern Kyushu and neighbouring areas; similar urns and coffins also are found on the Korean peninsula, where they probably originated. The graves were usually marked by mounds of earth or circles of stones, but a special type employed a dolmen (a large slab of stone supported over the grave by a number of smaller stones). Since the erection of dolmens was widely practiced in Manchuria and Korea, these, too, are believed to be a sign of an influx of continental culture. Normally, graves occur in clusters, but occasionally one is found apart, surrounded by a ditch and with swords, beads, and mirrors buried along with the dead. Such special graves suggest that society was already divided into classes.

While these new cultural elements represent a migration to Japan from the Korean peninsula or China, the migration was not of a magnitude to change the character of the people who had inhabited the islands from Jōmon times. Yayoi culture undoubtedly represents an admixture of new sanguineous elements, but it seems likely that the chief strain of proto-Japanese found throughout the country during the Jōmon period was not disrupted but was carried over into later ages. This point of view is supported by the accounts of the "people of Wo," found in the Chinese history *Wei chih.*

Chinese chronicles. Japan first appears in Chinese chronicles under the name of Wo (in Japanese, Wa). The Han histories relate that "in the seas off Lo-lang lie the people of Wo, who are divided into more than 100 states, and who bring tribute at fixed intervals." Lo-lang was one of the Han colonies established in the Korean peninsula.

The "state of Nu in Wo"

A history of the Later (Eastern) Han (AD 25–220) records that in AD 57 the "state of Nu in Wo" sent emissaries to the Later Han court and that the emperor gave them a gold seal. The "state of Nu," located around what is now Hakata Bay, in Kyushu, was one of more than 100 states that constituted Wo. In the latter half of the 2nd century, there was civil war in the state of Wo; Queen Himiko had pacified the land and, relying on her religious powers, ruled over a confederation of more than 30 states that maintained communications with the Wei dynasty (220–264) in China. Wei, too, sent emissaries to Wo, and friendly relations between the two sides continued during the first half of the 3rd century. The *Wei chih* contains a detailed account of the route from Lo-lang to the court of the Wo queen in "Yamatai."

Scholars are divided as to whether Yamatai was located in northern Kyushu or in the Kinai district (central Honshu). If it was in northern Kyushu, then the union of states was a purely local government, unrelated to the Yamato court of later times; but if it was in the Kinai district, then it would be natural to see it as the ancestor of that court. This would suggest, in turn, that Japan had already achieved a considerable degree of political unification. Japanese historians long sought to emphasize the antiquity and degree of unity of Yamatai in order to aggrandize Japan's relations with other East Asian nations. It seems most likely, however, that Yamatai was a local centre of power in Kyushu, and that further unification did not take place until at least a century later.

THE TUMULUS (TOMB) PERIOD (C. 250–552)

The unification of the nation. The questions of how the unification of Japan was first achieved and of how the Yamato court, with the *tennō* ("emperor of heaven") at its centre, came into being in central Honshu have inspired many hypotheses, none of which has so far proved entirely convincing. With the help of Chinese and Korean records, however, it is possible to get at least an approximate idea of the date by which substantial unification had occurred. The relations that Yamatai had begun with Wei were continued with the successor Chin dynasty (265–317); but, following the dispatch of a mission in 266, all records of exchanges cease, and it is not until 147 years later, in 413 during the Eastern Chin dynasty (317–419), that the name of Wo again appears in Chinese documents. The collapse of Yamatai and the birth pangs of the Yamato kingdom that took its place probably occurred during this period.

It is possible to push the date of unification of the nation back a few decades earlier than 413: a memorial erected in 414 commemorating the achievements of King Kwanggaet'o of Koguryŏ (a Korean state; 37 BC–AD 668), describing the fighting between Wo and Koguryŏ on the Korean peninsula from the end of the 4th century into the beginning of the 5th century, makes special mention of a great army sent to the peninsula in 391 by Wo. Such military success presupposes a long period of preparation. The 8th-century *Nihon shoki* ("Chronicles of Japan"), one of Japan's two oldest histories, mentions the dispatch of troops by Japan in 369. Such displays of strength would hardly have been possible unless Japan were already significantly unified, and the date of the unification of the country may therefore be about the mid-4th century at the latest. (T.Sa./G.C.H.III)

The Yamato court. Post-World War II historians have greatly revised the view of the place of Yamato in Asian affairs, downplaying the degree of control the Japanese formerly asserted that Yamato held over the Korean peninsula in ancient times. Most divide this period into three stages: a time of growth and expansion from about 250 to the end of the 4th century, a period of florescence that covers the 5th century, and then a period of decline from the early 6th century.

Three stages of the period

Rise and expansion of Yamato. The period is commonly called the Tumulus, or Tomb, period from the presence of large burial mounds (*kofun*), its most common archaeological feature. Whereas Jōmon and Yayoi burial practices were rather primitive, from the 3rd century large tombs, both circular and uniquely keystone-shaped, began to proliferate throughout Japan, marked most especially by the enormous tumuli in and around the Ōsaka area. It is from the very construction of the tombs themselves, from an examination of the grave goods, as well as from increasingly reliable written sources both domestic and foreign that a picture of the Yamato kingdom has emerged.

In the first stage of Yamato development, tombs clustered around the Shiki area of Yamato province (modern Nara prefecture), in the southwestern corner of the Nara

(Yamato) Basin. Rulers there held sway over an expanding portion of the archipelago. The Yamato kings were centred around Mount Miwa, the object of worship. Although they exercised both secular and sacred functions, it seems that their primary focus was a priestly one, based on a sacred connection with Mount Miwa. Archaeological findings suggest, however, that improved agricultural techniques—such as the use of iron tools for cultivation and improved techniques for leveling and flooding paddy fields—allowed the Yamato rulers to exercise control over significant manpower resources.

From about 350, power shifted north to the Saki area, near the present city of Nara. The nature of the burial goods in the tombs constructed there, the legendary accounts in *Kojiki* and *Nihon shoki,* as well as records from the continent all indicate that this was a period of Yamato expansion throughout the archipelago and even into the Korean peninsula. Although the rulers continued to worship Mount Miwa, the religious focus of the court seems to have been concentrated upon the Isonokami Shrine at Tenri, south of Nara. The rulers there seem to have been somewhat more military in nature than their Miwa predecessors, and archaeological findings suggest that the most treasured items of the Isonokami Shrine were in fact weapons—especially the so-called "seven-pronged sword" (*shichishitō*), which now is designated a National Treasure.

Thus, by the end of the 4th century, Yamato was in contact with Chinese rulers, exchanged diplomatic envoys with several of the kingdoms on the Korean peninsula, and was even strong enough to have sent an army against the powerful state of Koguryŏ, which then dominated the peninsula. Yamato was most closely associated with the southeastern kingdom of Paekche, whence came the "seven-pronged sword." Contact with the mainland, although involving conflict, also encouraged a marked rise in standards of living in the archipelago, as many of the fruits of advanced Chinese civilization reached Japan via people from the peninsula. Weavers, smiths, and irrigation experts migrated to Japan, and the Chinese ideographic script also was introduced at that time, together with Confucian works written in this script.

The Yamato court reached its peak in the early 5th century, during the second stage of its existence. Once again, there was a shift in the centre of power, this time directly westward to the provinces of Kawachi and Izumi (modern Ōsaka urban prefecture). The 5th century was one of spectacular development for Yamato, as evidenced by the enormous keyhole-shaped tombs in the suburbs of the modern Ōsaka region. It is now customary to regard the 5th-century rulers as a new line, distinct from those of the Shiki and Saki areas.

Tomb of Ōjin

What distinguishes the 5th-century tombs from earlier ones is both their enormous size—the tomb attributed to the semilegendary emperor Ōjin is some 1,380 feet (420 metres) in length—as well as their character. These rulers had access to great power in order to construct their tombs. It has been estimated that the construction of Ōjin's tomb would have taken 1,000 labourers, working from morning to night, four years to complete. The goods associated with these tombs are far more military in nature than those found in the earlier tombs: iron swords, arrowheads, and tools; armour; and all the trappings of a mounted warrior culture. All this suggests that the 5th-century rulers represent a more military, secular line of leaders in comparison with the priestly kings of the earlier Yamato area.

While most historians regard the 5th-century rulers as representing a new line, there is disagreement over their origin. Some have postulated an invasion of continental "horse riders" who seized control in the archipelago and established a new line of rulers. The consensus, however, still supports an indigenous shift in leaders relying on control of increased agricultural output and monopolizing superior military technology. From the court at Yamato, its rulers extended control along the Inland Sea and beyond, developing more sophisticated offices and units to control the peoples of the archipelago.

The Yamato polity. The pattern of administrative control established is called the *uji-kabane* system. *Uji* is usually translated as "clan" in English. The *uji* are thought to be extensions of original agricultural communities, perhaps what early Chinese records referred to as "states." Heads of the community functioned primarily as priests, mediating the relationship between the group and its deity. As clans joined together—probably largely by conquest—vertical relationships began to develop between heads of the communities and the queen or king at emergent courts. By the 5th century, these groups, possibly already called *uji,* were drawn together into economic, military, religious and familial ties with the Yamato kings. Some scholars have even argued that *uji* were purely political units, so designated by the Yamato ruler. *Uji* appeared first in the Nara Basin, in close association with the court; as the Yamato kingdom developed greater power, *uji* appeared in other areas as well.

By the 5th century, the Yamato ruler was designating the heads of the most powerful *uji,* who developed close ties with the ruler over time. The Yamato court was thus headed by a hereditary ruler, while its members were drawn from the group of powerful clan leaders awarded *kabane* (titles). The two major titles appear to have been *muraji* and *omi,* held only by clan leaders of powerful communities serving in the area of the Yamato court.

The *be*

Another factor that aided the expansion of the emergent state was the economic and military support of occupational groups, called *be* or *tomo,* attached to the court and its supporting *uji.* Structurally somewhat similar to clans, these occupational groups were distinguished by providing a special service to the court or to a superior clan. Earlier *be* were more likely to provide personal services or specialize in religious functions; but as the power of the Yamato court spread throughout the archipelago in the 5th century, newer *be* came to be involved with the production of weapons, armour, and mirrors or with the construction of irrigation systems. Some *be* were directly controlled by the court, including special ones called *nashiro* and *koshiro* set up for the support of certain royal relatives. Others were controlled by powerful clans directly in the service of the court.

Yamato relations with Korean states. If the 5th century represents an expansion of power throughout the archipelago, it also was a time of involvement in Korean affairs, as the struggle for peninsular hegemony intensified. At the time of Yamato's expedition against Koguryŏ in the late 4th century, Paekche and Yamato found themselves allied against Silla or Koguryŏ (or both); while the latter looked to northern Chinese kingdoms for support and legitimation, Yamato and Paekche usually turned to southern China. Paekche was frequently attacked by Koguryŏ during the century, prompting continued requests for assistance from Yamato; it is recorded that Paekche even sent a crown prince to Yamato as a hostage on one occasion and the mother of the king on another. Yet, probably because of internal dissension, Yamato did not dispatch any troops to the peninsula.

Yamato's interest in Korea was apparently a desire for access to improved continental technology and resources, especially iron. Yamato apparently gained a modicum of power in this region, controlled by the league of the Kaya (Japanese: Mimana) states between Paekche and Silla, though the exact relationship—whether ally or tributary—is unclear. But in the 6th century, Silla became militarily powerful, and Yamato was driven entirely from the peninsula when Silla annexed the Kaya league in 562.

Yamato decline and the introduction of Buddhism. The 6th century, in fact, represented a decline of Yamato power both at home and abroad. It was also marked by another shift of the court, this time back to the old region around Mount Miwa sometime late in the reign of Keitai (507–*c.* 531). From Keitai's reign there was a marked reduction in royal power. A campaign on the continent ended in defeat, signaling the decline of Yamato power. The rest of the 6th century can be characterized by the growing accumulation of power by regional clan leaders and a weakening of royal power, as well as the rise of new clans, mostly of recent continental origin.

The Soga clan

Chief among them were the Soga, who under the successive chieftains Iname and his son Umako rose to positions of dominance at court. Possessed of administrative and technical skills, especially in the fiscal area, the Soga established marriage connections with the royal house that permitted them considerable influence at court. The Soga are also known as sponsors of Buddhism at the Yamato court. Ultimately, the Soga clan eclipsed all other clans at court, especially after the destruction of the Mononobe clan in 587, and dominated the political scene. By the end of the 6th century, Japan had reached a low point in both foreign and domestic affairs. (G.C.H.III)

During the declining years of the Yamato court, however, there was one event of the utmost cultural importance: the introduction of Buddhism from Paekche. The date of its introduction is traditionally set at either 538 or 552, but it seems likely that Buddhist beliefs had begun spreading among the Japanese at a much earlier date. Buddhism at first was an object of wonder and admiration, a rare item of foreign culture symbolized by its beautiful statuary, its imposing religious paraphernalia, and its majestic temples. It was regarded as especially important in protecting the state. A true understanding of its doctrines did not come until the time of Prince Shōtoku (Shōtoku Taishi).

THE AGE OF REFORM (552–710)

The idealized government of Prince Shōtoku. The Yamato court was resuscitated by efforts made within the royal family itself, efforts that in the course of a century reformed the government of the country and set it moving toward formation of a centralized state more suited to the new age. This era is sometimes called the Asuka period for the region south of modern Nara where the royal courts were located. The movement was touched off by the theories of ideal government expounded by Prince Shōtoku. Shōtoku served as regent for his aunt, the empress Suiko (ruled 592–628), who was enthroned after the murder of her predecessor, Sushun (it was during Suiko's reign that the term *tennō,* or emperor, was adopted). Shōtoku took the Buddhist principles of peace and salvation for all beings as the ideal underlying his government. The prince's political policies, however, were based squarely on Chinese Confucian ideals.

The constitution of 604

The prince's most striking domestic achievements were the establishment of a system of 12 court ranks in 603 and the Seventeen-Article Constitution in 604. The former, which made clear the relative stations of court officials by giving them caps of different colours, aimed to encourage the appointment of men of ability and give the court a proper organization and etiquette of its own.

The constitution set forth the ideals of the state and rules for human conduct. It distinguished the ruler, government ministers, and the people as the three human elements making up the state and clearly laid down the duties and rights of each; it thus established the ideal of a centralized state presided over by a single ruler, and it provided a kind of basic law of the nation. The document not only shows the influence of Buddhism—of which the prince can be counted as the first major propagator in Japan—but it also embodies many of the ethical and political doctrines of Confucian government, long since established in China and subsequently implemented in the kingdoms on the Korean peninsula as well. By borrowing the ideas and vocabulary of continental government, Shōtoku attempted to buttress the legitimacy of the royal house, which had suffered diminution at the hands of great clans.

Shōtoku's chief achievement in foreign relations was the opening of relations with the Sui dynasty (581–618) of China. The exchanges between Japan and China in the 5th century had placed Japan in the position of a tributary state. Prince Shōtoku opened relations with Sui on an equal basis. Envoys were exchanged by the two countries. He also sent Japanese students to China to learn directly from Chinese culture, which had hitherto reached Japan via the states of Korea. Shōtoku was a profound student of Buddhism who gave lectures on the scriptures and himself wrote commentaries.

As Buddhism gained ground, imposing temples were built in the Chinese style. A new civilization descended on Japan almost overnight. Of the temples built at the time, all that has survived of most of them are the foundation stones, but the Hōryū Temple, founded between 601 and 607 at Ikaruga in present Nara prefecture, still preserves its ancient wooden structures; its extant buildings, dating from the late 7th and early 8th centuries, are the oldest wooden structures in the world.

The Taika reforms. The death of Prince Shōtoku in 622 prevented his ideals of government from bearing full fruit. The Soga family, regaining its former powers, killed Shōtoku's son Yamashiro Ōe and all his family in 643. At the same time, however, the students whom Shōtoku had sent to China were returning to Japan with accounts of the power and efficiency of the T'ang dynasty (618–907), which had overthrown the Sui dynasty and unified China. These accounts impressed on educated men the need to reform the government, strengthen the power of the state, and take every step to prepare against possible pressure from outside.

The coup d'état of 645

In 645 Prince Nakano Ōe and Nakatomi Kamatari engineered a coup d'ětat within the palace, killing the Soga family and wiping out all forces opposed to the imperial family. They then set about establishing a system of centralized government with the emperor as absolute monarch at its head. An edict issued in 646 abolished private ownership of land and people by powerful *uji.* The land thus taken over by the state was to be allocated among all who had attained a certain age, with the right to cultivate, in exchange for which the tenants were to pay a fixed tax. Provisions also were made for a governmental system embracing a capital city and local administration and for defense and communications facilities. The main outlines of the reforms were drawn up in about five years. They are given the name Taika reforms for the *nengō* ("year name")—the first such in Japanese history—that was given to the era at that time.

Not long after the Taika reforms, Japan did, in fact, become involved in a dispute that led it to again send troops to Korea. Paekche, whose capital fell in 660 to the combined forces of T'ang (China) and Silla, called on Japan for help. Japan, which had traditionally been friendly with Paekche, sent a large army; it was crushed, however, in 663, by a T'ang-Silla army at the mouth of the Kum River. Japan withdrew entirely and gave up any further intervention on the Korean peninsula.

Prince Nakano Ōe, ascending the throne as the emperor Tenji, directed his attention to domestic affairs. He built fortifications in Kyushu to prepare for an expected T'ang and Silla invasion and amended the system established by the Taika reforms so as to make it more suitable to the practical needs of the state. Upon Tenji's death, a fierce succession dispute erupted into warfare between the supporters of his younger brother and those of his uncle. His younger brother was victorious, and, as the emperor Temmu, upgraded the status of the Shintō shrine at Ise, making it the fountainhead of the dynasty's legitimacy; propagated Buddhism nationwide as a means of protecting and strengthening the state; ordered the compilation of official histories to enhance the prestige of the nation and, consequently, the dynasty; and had the Taika reforms codified as the Asuka Kiyomihara Code, from which the *ritsuryō* political structure emerged.

The *ritsuryō* system. The *ritsuryō* system refers to the governmental structure defined by *ritsu,* the criminal code, and *ryō,* the administrative and civil codes. Such a system had long been in force in China, and the Japanese *ritsuryō* was an imitation of the *lü-ling* of T'ang China and incorporated many of its original articles. The features were first delineated in rough form in the Taika edicts, but then were refined—perhaps first by Tenji in the Omi Code and then by Temmu—and certainly given final form in the Taihō Code of 701 and its successor, the Yōrō Code of 718.

Imperial government

Under the *ritsuryō* system, the Japanese emperor, for example, was in some respects an absolute monarch who ruled over the whole country as the head of a bureaucracy in the same manner as the emperor of China. Yet at the same time, he was also the traditional high priest who maintained peace for the land and people by paying tribute to the deities and sounding out their will. Thus, the central government was headed by twin agencies—the Council of State (Dajōkan), which combined within its functions the various practical aspects of administration, and the Office of Deities (Jingikan), a parallel bureaucracy for the worship of the deities. Prospective bureaucrats were required to study at a central college and to pass prescribed examinations. The recruitment of officials via examination was based on the highly developed bureaucratic system of China, yet the *ritsuryō* system was not too bound by its provisions to provide special favours for men of high rank and good family. This, too, was a compromise between the new principles of the *ritsuryō* system and the old spirit of respect for birth. In fact, the examinations were soon dropped. The provinces were divided into three types of administrative division: the *kuni,* or *koku* (province), the *kōri,* or *gun* (county), and the *sato,* or *ri* (village), to be administered by officials known as *kokushi, gunji,* and *richō,* respectively.

The people were divided into two main classes, freemen and slaves. The slaves were the possession of the government, the aristocracy, and the shrines and temples; as such, they were obliged to provide unlimited labour, but their total number accounted for less than one-tenth of the population. The majority of the free population were farmers. At the age of six, each male child was apportioned paddy fields that remained his to cultivate for life. A tax was levied on the produce of the paddies, and a head tax was levied on adult males. The paddy field tax was low, but the head tax, payable in handicrafts such as silk and hemp, imposed a heavy burden. Moreover, the transport of the goods from the provinces to the capital was the responsibility of the taxed, which involved an enormous labour for those living in distant parts. Adult males were also obliged to give military service and to provide labour for public works at the command of the local *kokushi.* Since the government's finances depended on such tribute from the common people, whenever the latter found the burden too much and fled to avoid paying taxes, government revenues quickly declined.

All land was, in principle, the property of the state. Most of the land was distributed equally among the people, but, apart from this, land of a certain annual yield was given to bureaucrats and other high-ranking persons as stipends and to Shintō shrines and Buddhist temples as sources of revenue. Land other than paddy fields was left to individuals to use as they pleased. There was a need to open up new paddy fields as a means of providing for a growing population, but the *ritsuryō* system made inadequate provision for this process. In fact, the complex taxation and allotment system discouraged the heavy investment necessary to open new paddy fields. Ultimately, the government had to encourage the opening up of new land by offering incentives; and in 743 the law was changed to allow permanent private possession of land to the person who had first put it under cultivation. As a result, the aristocrats and the shrines and temples set about putting land under cultivation in order to increase their own privately owned territories. The principle of public ownership of land provided for in the *ritsuryō* system began to crumble, and as it did so the whole system of government grew increasingly shaky.

Private ownership of land

THE NARA PERIOD (710–784)

Beginning of the imperial state. In 710 the imperial capital was shifted a short distance from Asuka to Nara. For the next 75 years, with minor gaps, Nara was the seat of government, and the old custom of changing the capital with each successive emperor was finally discarded. During this period, the centralized government provided for under the *ritsuryō* structure worked reasonably well; it was a time of atypical social mobility based on merit, where those with Chinese learning or Buddhist knowledge enjoyed access to power. Perhaps the most conspicuous feature is the brilliant flowering of culture, especially Buddhist culture. The leaders in its promotion were the emperor Shōmu and his consort, Kōmyō. Immediately on his accession, Shōmu—who from childhood had been given a thorough schooling as future emperor—showed an eager concern to promote the stable livelihood of the people. Convinced that the Buddhist faith was a means to ensure both the happiness of the individual and peace for the country as a whole, he introduced strong doses of Buddhism into his government.

One of the measures he took was the founding of the provincial temples known as *kokubunji.* Each province was to build a monastery (*kokubunji*) and a nunnery (*kokubun niji*), each with a seven-story pagoda and each housing a statue of the Śakyamuni Buddha. The second measure taken by Shōmu was the construction of the Tōdai Temple as *kokubunji* of the capital and the installation within it of a huge bronze figure of the Vairocana Buddha as supreme guardian deity of the nation. The great image that was produced as a result, though damaged in later ages, still stands in the Tōdai Temple and is famous the world over as the Great Buddha (Daibutsu) of Nara. The court also tried to attract Chinese monks to Nara. The most important of these was Ganjin (Chinese: Chien-chen), who finally reached Nara in 753 on his sixth attempt and founded the Ritsu sect at Tōshōdai Temple.

The marriage of Buddhism and politics that was Shōmu's ideal was to cause trouble after his death. The temples gradually amassed vast wealth, and the monks acquired high political positions and began to interfere in secular affairs. A movement to counter such abuses arose among the aristocracy, the leaders of the movement being the Fujiwara family, descendants of Nakatomi Kamatari, who had played such an important role in the Taika reforms. Kamatari and his son Fuhito (both later given the surname Fujiwara) had supervised compilation of the Taihō and Yōrō codes that formalized the *ritsuryō* system and had become prominent figures at court as a new type of bureaucrat-noble. Moreover, Shōmu's marriage to Fuhito's second daughter (who became known as the empress Kōmyō) created the precedent for a marital relationship with the imperial house that was to last throughout much of premodern Japanese history. The subsequent progress of the family's fortunes in the Nara period was not always smooth, however.

Emergence of the Fujiwara

In particular, the emphasis on Buddhism undercut the family's influence. At the end of the 8th century, the powerful priest-premier Dōkyō rose to a position of undisputed hegemony under Shōmu's daughter, who reigned twice, as the empress Kōken and then as Shōtoku; and Fujiwara nobles feared that the priestly domination of government threatened the future of the nation. Ousting Dōkyō following the death of the empress, they set on the throne a new emperor, Kōnin, who was less enthralled with Buddhism. Kōnin's son, the emperor Kammu, who was of a similar mind, shifted the capital first to Nagaoka and in 794 to Heian (or Heian-kyō; present Kyōto) to sever connections with the temples of Nara and reestablished government in accordance with the *ritsuryō* system. Kammu's accession also represented a shift from the descendants of the emperor Temmu back to those of Tenji, whose base of power was located in Yamashiro province, the site of the new capital.

Culture in the Nara period. The cultural flowering centring on Buddhism was an outcome of lively exchanges with other nations. Four times within 70 years the government sent official missions to the T'ang court, each mission accompanied by a large number of students who went to study in China. By this time T'ang had formed a great empire that controlled not only the central plains of China but parts of Mongolia and Siberia to the north and of Central Asia to the west.

Nara culture, borrowing from the T'ang, whose capital, Ch'ang-an, was a great international city, evinced a marked international flavour itself. But despite this internationalism, respect was also shown for traditional Japanese cultural forms. An outstanding example of this respect is the collection of Japanese verse known as *Man'yōshū* (*c.* 8th century AD), an anthology of 4,500 poems both ancient and contemporary. Poets represented in the anthology range over all classes of society, and the scenery celebrated in the verse represents districts throughout the country. The anthology had immense influence on all subsequent Japanese culture. The compilation of Japan's two most ancient histories, the *Kojiki* and *Nihon shoki,* also took place at the beginning of the 8th century. The histories—a combination of myth, folk belief, and, as they near the

The *Kojiki* and *Nihon shoki*

contemporary age, historical fact—were highly political in nature: by stressing the connection between the imperial family and the sun goddess (Amaterasu), they provided a written legitimation of the rule of the imperial house.

THE HEIAN PERIOD (794–1185)

Changes in *ritsuryō* government. In 794, as noted above, the emperor Kammu shifted his capital to Heian, diluted the ties between government and Buddhism, and attempted to revive government in accordance with the *ritsuryō*. Commanding that the provisions of the *ritsuryō* system be enforced, he also amended those articles that were no longer relevant to the age. Since it was difficult in practice to carry out the allocation of rice fields once every 6 years, this was amended to once in 12 years. A tighter watch was imposed on corruption among local officials. The original system of raising conscript troops from among the peasantry was abolished, and soldiers were thenceforth selected from among the sons of local officials with martial prowess. Kammu dispatched large conscript armies against the Ezo (Emishi), a nonsubject tribal group in the northern districts of Honshu who were regarded as aliens. The Ezo eventually were pacified, although the northern border was never fully brought under the control of the central government.

Interference in affairs of state by religious authorities was forbidden, but they were encouraged to see that Buddhism fulfilled its proper functions. Kammu was a supporter of Buddhism for both national and individual purposes. He dispatched two brilliant monks, Saichō and Kūkai, to China to study. Each of them, on his return to Japan, established a new sect of Japanese Buddhism: the Tendai sect, founded by Saichō, and the Shingon sect, established by Kūkai. The two sects were thenceforth to form the mainstream of Japanese Buddhism.

After Kammu, successive emperors carried on his policies, and society enjoyed some 150 years of peace. The formal aspects of government, at least, were carefully observed. The social reality, however, became increasingly chaotic, so that form and actuality were soon traveling along quite different courses. The very foundations of *ritsuryō* government began to crumble because of the difficulty of carrying out the allotment system based on census registers and the consequent decline in government revenue. Two changes were instituted early in the 10th century that, while temporarily shoring up government finances, eventually led to further erosion of the ideals of the authority-intensive *ritsuryō* system. First, the state decided to calculate taxes on the basis of land units rather than individuals. The government set up taxation units based on paddy fields upon which both rent and corvée could easily be assessed. Second, the central government gave up the details of administering provincial affairs, leaving local matters to governors (now increasingly called *zuryō*, or "tax managers") and local resident officials (*zaichō kanjin*) who were mainly responsible for forwarding to Heian a specified tax amount. It now became easier to calculate the amount of taxable public land (*kōden*) in each province; but entrusting so much authority to governors opened the gates for further abuse, especially the possibilities of increasing the amount of lands held in tax-free estates.

Rise of local officials

Another example of the divergence between form and reality is the fact that while, on the surface, appointments to official posts were made in accord with *ritsuryō* stipulations, real power shifted to other posts that were newly created outside the codes as the occasion demanded. Early examples were the two new posts created during the early 9th century: *kurōdo,* a kind of secretary and archivist to the emperor, and *kebiishi,* the imperial police, who ultimately developed powers to investigate crimes and determine punishments. The two most important posts developed outside the *ritsuryō* codes were those of *sesshō* (regent) and *kampaku* (chief councillor), better known by an abbreviated combination of the two terms, *sekkan* (regency). The original role of the *sesshō* was to attend to affairs of state during the minority of the emperor, while the *kampaku*'s role was to attend to state matters for the emperor even after he had come of age. Neither post had been foreseen by the *ritsuryō* system, which was based on the principle of direct rule by the emperor.

In the mid-9th century, however, when nine-year-old Seiwa ascended the throne, his maternal grandfather, Fujiwara Yoshifusa, created the office of *sesshō,* based on the post once held by imperial family members such as the empress Jingū and the princes Nakano Ōe and Shōtoku. Yoshifusa's son Mototsune became *sesshō* during the minority of the succeeding emperor Yōzei, and then in the reign of the emperor Uda he created the post of *kampaku.* It thus became the established custom that a member of the Fujiwara family should serve as *sesshō* and *kampaku.* In order to hold the *sekkan* offices, it was necessary that the person concerned should marry his daughter into the imperial family, then establish the resulting offspring as emperor. While not totally new with the Fujiwara—the maternal relatives of the early Yamato rulers (notably the Soga) were the important powers at court—the system reached its height and perfection under the Fujiwara. As a result of this complex system, there were constant struggles at court involving the expulsion of members of other families by the Fujiwara family or wrangling among the branches of extensive Fujiwara clan itself.

The culture of the 9th century was a continuation of that of the 8th, insofar as its foundations were predominantly Chinese. The writing of Chinese prose and verse was popular among scholars, and great respect for Chinese customs was shown in the daily lives of the aristocracy. Buddhist monks continued to travel to China to bring back as yet unknown scriptures and iconographic pictures. Buddhist sculpture and paintings produced in Japan were done in the T'ang style. At the end of the 9th century, however, Japan cut off formal relations with T'ang China, both because of the expense involved in sending regular envoys and because of the political unrest accompanying the breakup of the T'ang empire. The practical result was the stimulation of a more purely Japanese cultural tradition. Japanese touches were gradually added to the basically T'ang styles, and a new culture slowly came into being; but it was not until the 10th century and later that this tendency became a strong current.

Aristocratic government at its peak. From the 10th century and through the 11th, successive generations of the northern branch of the Fujiwara clan continued to control the nation's government by monopolizing the posts of *sesshō* and *kampaku,* and the wealth that poured into their coffers enabled them to lead lives of the greatest brilliance. The high-water mark was reached in the time of Fujiwara Michinaga (966–1028). Four of his daughters became consorts of four successive emperors, and three of their sons became emperors. Government during this period was based mostly on precedent, and the court had become little more than a centre for highly ritualized ceremonies.

The *ritsuryō* system of public ownership of land and people survived in name alone; land passed into private hands, and people became private citizens. The fiscal changes of the early 10th century did not bring enough paddy fields into production, and tax rates remained high. Public revenue—the income of the Heian aristocrats—continued to decline, and the incentive to seek new private lands increased. Privately owned lands were known as *shōen* ("manors"), which developed primarily on the basis of rice fields under cultivation since the adoption of the *ritsuryō* system. Since the government-encouraged opening up of new land during the Nara period, temples and aristocrats with resources at their disposal had hastened to develop new areas, and vast private lands had accrued to them. Originally private lands had been taxable, but *shōen* owners developed various techniques to obtain special exemption from taxes, so that by mid-Heian times the *shōen* gradually became nontaxable estates. The increase in *shōen* thus came to pose a serious threat to the government, which accordingly issued edicts intended to check the formation of new estates. This merely served, however, to establish more firmly the position of those already existing and failed to halt the tendency for such land to increase. Finally, an edict issued in 1069 recognized all estates established before 1045.

The growth of *shōen*

Although the aristocracy and temples around the capital enjoyed exemption from taxes on their private lands, the same privileges were not available to powerful families in the provinces. These, accordingly, commended their holdings to members of the imperial family or the aristocracy, concluding agreements with them that the latter should become owners in name while the former retained rights as actual administrators of the property. Thanks to such agreements, the estates of the aristocracy increased steadily, and their incomes swelled proportionately. The *shōen* of the Fujiwara family expanded greatly, especially in the 11th and 12th centuries.

While the aristocracy was leading a life of luxury on the proceeds from its estates, the first stirrings of a new power in the land—the warrior, or samurai, class—were taking place in the provinces. Younger members of the imperial family and lower-ranking aristocrats dissatisfied with the Fujiwara monopoly of high government offices would take up posts as local officials in the provinces, where they settled permanently, acquired lands of their own, and established their own power. In order to protect their territories or expand their power, they began to organize local inhabitants (especially the *zaichō kanjin*) into service. A number of powerful provincial aristocrats developed significant armed forces. As a consequence, when men of true martial ability and sufficient autonomy emerged, the slightest incident involving any one of them might provoke armed conflict. The risings of Taira Masakado (d. 940) in the Kantō district and of Fujiwara Sumitomo (d. 941) in western Japan are examples of large war bands extending their control in the provinces. Although the government was able to suppress the rebellions, these conflicts had an enormous effect in lowering the government's prestige and encouraging the desolation of the provinces.

During the 10th century a truly Japanese culture developed, one of the most important contributing factors being the emergence of indigenous scripts, the *kana* syllabaries. Until then, Japan had no writing of its own; Chinese ideographs were used both for their meaning and for their pronunciation in order to represent the Japanese language, which was entirely different grammatically from Chinese. Educated men and women of the day, however, gradually evolved a system of writing that used a purely phonetic, syllabic script formed by simplifying a certain number of the Chinese characters; another script was created by abbreviating Chinese characters. These two scripts, called *hiragana* and *katakana*, respectively, made it possible to write the national language with complete freedom, and their invention was an epoch-making event in the history of the expression of ideas in Japan.

The development of the *kana* syllabaries

Particularly noteworthy in this respect were the daughters of the Fujiwara, who, under the aristocratic government of the day, became the consorts of successive emperors and surrounded themselves with talented women who vied with each other in learning and the ability to produce fine writing. The *hiragana* script—largely shunned by men, who composed official documents in stilted Chinese—provided such women with an opportunity to create works of literature. Among such works, *The Tale of Genji* (*Genji monogatari*), by Murasaki Shikibu, and *The Pillow Book of Sei Shōnagon* (*Makura no sōshi*), by Sei Shōnagon, are masterpieces of world literature.

By Heian times, the diverse poetic forms found in the *Man'yōshū* had been refined into one form called waka. The waka, consisting of 31 syllables, was an indispensable part of the daily lives of the aristocracy, and proficiency in verse-making was counted an essential accomplishment for a courtier. The value placed on the skillful composition of poetry led to the compilation in 905 of the *Kokinshū* (or *Kokin wakashū*), the first of a series of anthologies of verse made at imperial command.

The same trend toward the development of purely Japanese qualities became strongly marked in Buddhism as well. Both the Tendai and Shingon sects continued to flourish. But they tended to pursue worldly wealth and riches at the expense of purely religious goals, and it was left to the Pure Land (Jōdo) sect of Buddhism to preach a religion that sought to arouse a desire for salvation in ordinary people. Pure Land Buddhism expounded the glories of the paradise of Amida (Amitābha, or Buddha of Infinite Light)—the world after death—and urged all to renounce the defilements of the present world for the sake of rebirth in that paradise; it seemed to offer an ideal hope of salvation in the midst of the disorder and decay of the old order. It grew in popularity as society began to unravel and violence spread at the end of the Heian period. Pure Land eschewed difficult theories and ascetic practices, teaching that in order to achieve rebirth it was necessary only to invoke the name of Amida. This same teaching also inspired artists to produce an astonishing number of representations of Amida in both sculpture and painting.

Another example of this Japanization of culture is the style called Yamato-e ("Japanese painting"). Most Yamato-e dealt with secular affairs—for example, *The Tale of Genji*—and there were even satirical works lampooning the behaviour of the court nobles. The signs of the growing independence of Japanese culture, apparent in every field, were an indication that, by now, two centuries after the first ingestion of continental culture, the process of naturalization was nearing completion.

Yamato-e style

Government by cloistered emperors. The powerful authority wielded by the Fujiwara regents was maintained by their maternal relationship to successive emperors; once such a relationship disappeared, their power was bound to weaken. This is, in fact, what happened in late Heian times. The emperor Go-Sanjō ascended the throne in 1068, the first sovereign in more than a century not born of a daughter of the Fujiwara; while Michinaga's sons Yorimichi and Norimichi both gave their daughters to be imperial consorts, no Fujiwara-related heirs resulted from these unions. As a result, the adult Go-Sanjō began to rule free of the strong control of a Fujiwara regent. His policies were designed both to strengthen the weakening economic institutions of the state and to bolster the fortunes of the imperial family itself.

After only four years on the throne, Go-Sanjō abdicated and, in accord with the precedent established by earlier emperors, opened an office of the retired emperor (*in no chō*). Since Go-Sanjō clearly meant to participate in politics even from retirement, especially to direct the imperial succession to his non-Fujiwara sons, his era is often regarded as the institutionalization of rule by retired or cloistered emperors.

Go-Sanjō died shortly after abdicating, but he was followed by three successive rulers—Shirakawa, Toba, and Go-Shirakawa—who exercised sovereign power both as emperors and then even more effectively as retired emperors. Governmental control in Japan thus passed from Fujiwara regents to the "cloistered emperors" who wielded real power behind the scenes during the late 11th and 12th centuries. This system, known as *insei* ("cloistered government") because the retired emperors all took Buddhist vows and retired to cloisters (*in*), was not dramatically different from the manner in which Fujiwara regents had ruled. Decisions continued to be made by a relatively small group of high-ranking nobles, the majority of whom were now clients of the retired emperor rather than the Fujiwara regent. The reigning emperor was largely treated as a figurehead; now, however, control over this position returned to the hands of imperial family, allowing it to compete more effectively for the rewards of power.

The cloistered emperor system continued for a long period, although the emperors Shirakawa, Toba, and Go-Shirakawa were the only ones to wield absolute, behind-the-scenes power. *Insei* represented a revival of imperial family fortunes: with a vibrant household organization, the ability to attract clients among the nobility, and the opportunity to attract *shōen* holdings of its own, the fortunes of the house increased immeasurably. By the end of the Heian period, in fact, the imperial family had eclipsed the Fujiwara as the largest *shōen* holder in the land.

One common feature of each reign was that the retired sovereign became a Buddhist priest and governed in a way that theoretically respected the teachings of Buddhism. In practice, however, retired emperors seemed more concerned with the construction of ostentatious temples; temples also were endowed with *shōen* commended by clients

Mappō

of the imperial family. The secularization of Buddhism continued apace. Late Heian times were the "latter days" (*mappō*) of Buddhist calculation, in which one could rely upon nothing but faith in some Buddhist deity or doctrine for salvation. In hopes of salvation, many aristocrats donated funds to construct temples or took holy vows and went to live in temples, which thus became centres of political intrigue. Most higher positions in the religious world were occupied by members of the imperial family and former aristocrats. This effectively closed advancement to commoners, and the lower-ranking monks in the temples often resented their superiors on this account. Whenever some particularly serious grievance arose, they would march in a body on the capital and try to force acceptance of their demands by a direct appeal to the court. Nor did the monks hesitate to resort to armed force; it was an age in which some members of a priesthood ostensibly committed to compassion and respect for life in all its forms could openly bear arms and engage in slaughter.

The rise of the warrior class. In the late Heian period, the more powerful of the samurai, who, as noted above, first established their power in the provinces, gradually gathered in or near the capital, where they served both the military needs of the state against potential outbreaks of rebellion and also as bodyguards for the great noble houses. Through association with the aristocracy, they gradually established a foothold at court. Outstanding among these samurai were the branch of the Minamoto (or Genji) family descended from the emperor Seiwa and the Taira (Heike) family lineage that traced its roots to the emperor Kammu. The Seiwa Genji established themselves as clients in the service of successive Fujiwara regents even before Michinaga was regent. Their fame as a warrior clan was greatly heightened in the mid-11th century when they quelled a rebellion in northeastern Japan. The victorious Minamoto leader Yoshiie became the nation's most celebrated warrior, and many local figures made voluntary vows of allegiance to him and commended lands to him in return for his protection. Yoshiie's sudden rise to power forced the court to view him warily, even denying the commendation of estates from would-be clients.

The Taira had at first settled in the Kantō district, where they extended their influence over a wide area; but they had suffered a setback with the defeat of Taira Masakado and had finally lost their hold in the Kantō district as the result of another later uprising by Masakado's descendant Tadatsune. With the revitalization of the imperial family, the Taira curried favour with the retired emperors. Taira Masamori and his son Tadamori served as governors in several western provinces, building up their own power in the area. Tadamori also initiated trade with Sung dynasty China as a means of amassing wealth. Because they were clients of the retired emperor, the social position of the Taira rose steadily, so that Tadamori's son Kiyomori broke into the ranks of the nobility.

Discord within both the imperial family and the Fujiwara regent's house split the nobility into two factions, each of which enlisted warriors from the Minamoto and the Taira. The two factions eventually clashed openly in Kyōto in what is known as the Hōgen Disturbance (1156). The conflict was highly significant in that it demonstrated the inability of the courtiers to settle major differences without reliance on the power of the warriors. Conflicts over rewards arose between the two successful Hōgen generals, Minamoto Yoshitomo and Taira Kiyomori, and, in the Heiji Disturbance (1159) that followed, the two warrior clans were pitted against one another. The Minamoto were thoroughly defeated, and Taira Kiyomori emerged as a major power in the land.

The Heiji Disturbance

Although Kiyomori was born into a middle-ranking provincial warrior family, he became, in effect, a military noble and dominated the political scene in ways reminiscent of the Fujiwara. Over the two decades following the Heiji Disturbance, Kiyomori and his kinsmen gradually assumed power at court. Kiyomori himself became prime minister (*dajō-daijin*), and many other official posts were filled by members of his family. All his daughters were married into powerful noble families, and one even became the consort of the emperor Takakura. The infant prince born of their union ascended to the throne in 1180 as the emperor Antoku, and Kiyomori's power rose even higher through his influence over the throne, which represented a return to government by matrilineal relatives of the emperor. Kiyomori's rule also had its more drastic aspects. In a single move, for example, he swept 42 court officials from their posts and into exile, and he razed to the ground such troublesome places as the Tōdai and Kōfuku temples. His repairing of the Inland Sea route, however, and his encouragement of trade with Sung China—by which the Taira became wealthy—were farseeing measures that distinguished Kiyomori from earlier Fujiwara regents.

The high-handed manner in which Kiyomori and his kinsmen dominated the court, however, naturally provoked reaction. While the Taira thrived in the capital, the descendants of the Minamoto quietly built up their strength in the provinces. Finally Yoritomo, the oldest surviving son of Yoshitomo, rallied the Minamoto and other great warrior families in eastern Japan in insurrection. From the initial uprising in 1180 to the final sea battle at Dannoura at the southernmost tip of Honshu, the so-called Gempei (Genji and Heike) War engulfed Japan in warfare on a scale theretofore unseen. Although traditionally portrayed as a simple Taira-versus-Minamoto conflict, the Gempei War was in actuality a combination of interclan and intraclan fighting, as well as a struggle between central control and forces for local autonomy combined under the larger banner of clan rivalry. The final rout of the fleeing Taira forces on the sea, however, put a more or less decisive end to the swing of fortune between Minamoto and Taira.

It also marked an important turning point in Japanese history, since Yoritomo's establishment of a military government (*bakufu,* or shogunate, as it is often called in English) in Kamakura may be seen as the commencement of rule by a samurai class and at least the beginning of the end of the ancient monarchical system of court and aristocracy. In one form or another, a *bakufu* (literally, "tent government," the name for the field headquarters of a campaigning warrior), was to hold effective political control in Japan until the restoration of imperial power in 1868. (T.Sa./G.C.H.III)

Medieval Japan

THE KAMAKURA PERIOD (1185–1333)

The establishment of warrior government. The establishment of the *bakufu* by Minamoto Yoritomo at the end of the 12th century can be regarded as the beginning of a new era, one in which independent government by the warrior class successfully opposed the political authority of the civil aristocracy. During the Kamakura period, total warrior dominance was not achieved. There was, instead, what approached a dyarchy with civil power in Kyōto and military power in Kamakura sharing authority for governing the nation. Institutions of the Heian imperial-aristocratic system remained in place throughout the Kamakura age, replaced with new feudal institutions when Kamakura passed from the scene.

Dyarchy of power

During the Gempei War, Yoritomo established his headquarters in Kamakura and entrusted the suppression of the Taira to his younger brothers Noriyori and Yoshitsune. Meanwhile, he gathered a following of great eastern warrior leaders and began to lay the foundation for a new military government (the term *bakufu* was used only later in retrospect).

In 1185, after the destruction of the Taira family at the Battle of Dannoura, Yoritomo was granted the right to appoint his vassals, or *gokenin* ("housemen") as military governors (*shugo*) in the provinces and military stewards (*jitō*) in both public and private landed estates. It was the job of the *shugo* to recruit metropolitan guards and keep strict control over subversives and criminals. The *jitō* collected taxes, supervised the management of landed estates, and maintained public order.

Although the Gempei War ended in 1185, a dispute between Yoritomo and his brother Yoshitsune resulted in continued warfare until 1189. Three years later Yoritomo went to Kyōto and was appointed shogun (an abbreviation of *seii taishōgun;* "barbarian-quelling generalissimo"), the highest honour that could be accorded a warrior. Though he kept the title only briefly and was not known by that term in the documents he issued to manage Kamakura affairs, "shogun" ultimately emerged as the title associated with the head of a *bakufu.* At first the chief base of the Kamakura *bakufu* lay in the *shōen* seized from the Taira family and in the limited administrative revenues from public estates in provinces granted to Yoritomo by the imperial court. But later the *bakufu* was able to expand its influence over lands that were still controlled by the civil provincial governors, as well as the private estates of the civil aristocracy and the temples and shrines.

The Hōjō regency. After the death of Yoritomo in 1199, real power in the *bakufu* passed into the hands of the Hōjō family, from which Yoritomo's wife, Masako, had come. In 1203 Hōjō Tokimasa, Masako's father, assumed the position of regent (*shikken*) for the shogun, an office that was held until 1333 by nine successive members of the Hōjō family. Taking advantage of disputes among Yoritomo's generals, the Hōjō overthrew and outmaneuvered their rivals, and after three generations the direct line of descent from Yoritomo had become extinct. Kujō Yoritsune, a Fujiwara scion and distant relative of Yoritomo, was appointed shogun, while Tokimasa's son Hōjō Yoshitoki (*shikken* 1205–24) handled most government business. Thereafter, the appointment and dismissal of the shogun followed the wishes of the Hōjō family. Shoguns were selected only from the Fujiwara or imperial houses, out of concern for pedigree.

The increasing political power of the military led to a conflict with the aristocracy. Hence, the emperor Go-Toba in 1221 issued a mandate to the country for the overthrow of Yoshitoki. Few warriors, however, responded to his call. Instead, the Hōjō family dispatched a *bakufu* army that occupied Kyōto, and Go-Toba was arrested and banished. This incident is known as the Jōkyū Disturbance, named for the era name Jōkyū (1219–22). The *bakufu* now set up a headquarters in Kyōto to supervise the court and to control the legal and administrative business of the western provinces. The estates of the civil aristocrats and warriors who had joined Go-Toba were confiscated, and Kamakura vassals were appointed to *jitō* posts in them as rewards. The political power of the *bakufu* now extended over the whole country.

The Jōkyū Disturbance

Meanwhile, the regent Hōjō Yasutoki, to strengthen the base of his political power, reorganized the council of leading retainers into a Council of State (Hyōjō-shū). In 1232 the council drew up a legal code known as the Jōei Formulary (Jōei Shikimoku). Its 51 articles set down in writing for the first time the legal precedents of the *bakufu.* Its purpose was simpler than that of the *ritsuryō,* the old legal and political system of the Nara and Heian civil aristocracy. In essence, it was a body of pragmatic law laid down for the proper conduct of the warriors in administering justice.

©1996, Encyclopædia Britannica, Inc.

Important Japanese historical sites.

The Mongol invasions. The establishment of the regency government coincided with the rise of the Mongols under Genghis Khan in Central Asia. Beginning in 1206, in the space of barely half a century, they had established an empire extending from the Korean peninsula in the east to as far west as Russia and Poland. In 1260 Genghis Khan's successor, Kublai, became Great Khan in China and fixed his capital at present-day Peking (Beijing). In the autumn of 1274 a Mongol and Korean army of some 40,000 men set out from present-day South Korea. On landing in Kyushu it occupied a portion of Hizen province (part of present-day Saga prefecture) and advanced to Chikuzen. The Kyushu military vassals were mobilized for defense. A Mongol army landed in Hakata Bay, forcing the Japanese defenders to retreat to Dazaifu; but a typhoon suddenly arose, destroying more than 200 ships of the invaders, and the survivors returned to southern Korea.

The *bakufu* took measures to better prepare for a renewed invasion. Coastal defenses were strengthened, and a stone wall was constructed extending for several miles around Hakata Bay to thwart the powerful Mongol cavalry. Apportioned among the Kyushu vassals, these public works took five years to complete and required considerable expenditure. Meanwhile, the Mongols made plans for a second expedition. In 1281 two separate armies were arrayed: an eastern army consisting of about 40,000 Mongol, northern Chinese, and Korean troops set out from South Korea, and a second army of about 100,000 troops from southern China under the command of the Mongol general Hung Ch'a-ch'iu. The combined assault breached the defenses at Hakata Bay. But again a fierce typhoon destroyed nearly all of the invading fleet, forcing Hung Ch'a-ch'iu to retreat precipitately. The remnants of the invading army were captured by the Japanese.

The defeat of the Mongol invasions was of crucial importance in Japanese history. The military expenditure on preparations, continuous vigil, and actual fighting undermined the economic stability of the Kamakura government and led to the insolvency of many of the *jitō*. The bond between the Hōjō and the Kamakura vassals was strained to the breaking point. The invasions also led to another prolonged period of isolation from China that was to last until the 14th century. Moreover, the victory gave a great impetus to a feeling of national pride, and the *kamikaze* ("divine wind") that destroyed the invading hosts gave the Japanese the belief that they were a divinely protected people.

The *kamikaze*

Samurai groups and farming villages. The Japanese feudal system began to take shape under the Kamakura *bakufu*, though it remained only inchoate during the Kamakura period. Warrior-landlords lived in farming villages and supervised peasant labour or themselves carried on agriculture, while the central civil aristocracy and the temples and shrines held huge public lands (*kokugaryō*) and private estates in various provinces and wielded power comparable to that of the *bakufu*. These *shōen* were managed by influential resident landlords who had become warriors. The leaders of a large number of villagers, *jitō*, laboured to develop the rice fields and irrigation works in the areas under their jurisdiction.

Among these landlords, some were vassals of the shogun, while others were connected to the aristocracy or the temples and shrines. The *jitō* owed their loyalty to the shogun. In return, the shogun not only guaranteed these men security of tenure in their traditional landholdings but rewarded them with new holdings in confiscated lands. This connection between lord and vassal, on which grants of landownership or management were based, gave Japanese society a somewhat feudal character.

But these lands were by no means complete fiefs: the Kamakura *bakufu* did not possess large tracts of its own land that it could grant to its vassals as fiefs in return for service. Kamakura warriors could control traditional land types (*shōen* and *kokugaryō*) or be newly appointed into confiscated lands. In either case, there was a nominal absentee central proprietor—temple, shrine, or aristocratic or royal family—who maintained substantial control over the land. Thus, there was a limit on the degree to which the Kamakura warrior could exploit the land and people under his control. Conflict was endemic between central proprietor (usually a local representative of the proprietor) and *jitō:* the former wished to maintain as much control and income as possible while the latter was concerned with expanding his share. Since the *jitō* was entirely under the control of Kamakura, disputes flooded the warrior headquarters from landowners seeking to curtail *jitō* encroachments. Thus, the primary focus of Kamakura activity became the dispensing of justice in legal cases involving land disputes. The Kamakura *bakufu* gained a reputation for fairness. By various means, however, Kamakura warriors managed to whittle away significantly the absentee control of *shōen* proprietors.

Conflict also was endemic between the farming population and the warriors. There were several different statuses among the peasantry, including *myōshu,* prominent farmers with taxable, named fields (*myōden*) of significant size and long standing; small cultivators with precarious and shifting tenures; and others who paid only labour services to the proprietor or *jitō*. The lowest peasant category, called *genin* ("low person"), was made up of people who were essentially household servants with no land rights.

The samurai, in theory, performed military service on the battlefield and during times of peace, in addition to managing agricultural holdings, engaging in hunting and training in the martial arts, and nourishing a rugged and practical character. There was an emerging sense of ideal warrior behaviour that grew out of daily training and the experience of actual warfare. Pride of family name was especially valued, and loyal service to one's overlord became the fundamental ethic. This was the origin of the more highly developed sense of a warrior code of later ages. The status of women in warrior families was comparatively high; like their Heian predecessors, they were allowed to inherit a portion of the estates and even *jitō* posts, a practice that gradually came to be restricted.

Ideal warrior behaviour

After the middle of the Kamakura period, the farming villages in which the warriors resided underwent changes as agricultural practices advanced; other aspects of society were changing as well. Artisans were frequently attached to the proprietors of the *shōen* and progressively became more specialized, responding to a specific growth of consumer demand. Centres for metal casting and metalworking, paper manufacture, and other skills appeared outside the capital, in various provincial localities, for the first time. The exchange of agricultural products, manufactured goods, and other products thrived; local markets, held on three fixed days a month, became common. In the large ports along the Inland Sea and Lake Biwa, specialized wholesale merchants (*toimaru*) appeared who, as contractors, stored, transported, and sold goods. Further, it became common for many merchants and artisans to form guilds, known as *za,* organized under the temples, shrines, or civil aristocrats, from whom they gained special monopoly privileges and exemptions from customs duties.

Kamakura culture: the new Buddhism and its influence. During the Kamakura period the newly arisen samurai class began to supercede the ancient civil aristocracy, which nonetheless continued to maintain the classical culture. Vigorous overseas trade expanded contacts with the continent, fostering the introduction of Zen Buddhism (in Chinese, Ch'an) and Neo-Confucianism from Sung China. Chinese influences could be seen in monochrome painting style (*suiboku-ga*), architecture, certain skills in pottery manufacture, and the custom of tea drinking.

In matters of religion, the great social changes that took place between the end of the Heian period and the early Kamakura period fostered a sense of crisis and religious awakening and caused the people to demand a simple standard of faith, in place of the complicated teachings and ceremonies of the ancient Buddhism. Several new Buddhist sects sprang up that eschewed difficult ascetic practices and recondite scholarship. Among these may be included the Jōdo, or Pure Land, sect mentioned earlier and its offshoot, the Shin (True) school, and the sect established by the former Tendai priest Nichiren. By contrast, the Zen school sought to open the way to insight by self-effort (*jiriki*); hence, it met with a ready response, satisfying the demands of many samurai. At the same time, scholarship and the arts were still deeply linked with the Tendai and Shingon sects of esoteric Buddhism, which was a vigorous influence even in Shintō circles. Nonetheless, the new forms of worship expanded popular participation in Buddhism tremendously.

Literature

In scholarly and literary circles, the Kyōto nobility confined themselves largely to the annotation and interpretation of the ancient classics and to the study of precedents and ceremonies. But at the beginning of the Kamakura period, a brilliant circle of waka poets around the retired emperor Go-Toba produced a new imperial selection of poems entitled the *Shin kokin wakashū*.

Just before the Jōkyū Disturbance the Tendai monk Jien (a member of the Fujiwara family) completed his *Gukanshō* ("Jottings of a Fool"). This is the first work of historical philosophy in Japan to incorporate a notion of historical causality. Meanwhile, as warriors began to contend and mingle with court nobles, many warrior leaders developed a love of scholarship and a delight in waka poetry. Reflecting the rise of the warrior class, military epics became popular. The most famous is the anonymously written *The Tale of the Heike* (*Heike monogatari*). After the middle Kamakura period, as Buddhist pessimism grew fainter, various kinds of instruction manuals and family injunctions were composed, while collections of essays such as Yoshida Kenkō's *Essays in Idleness* (*Tsurezuregusa*) also made their appearance.

In the visual arts the carving of wooden images of famous monks flourished, and, after the middle of the Kamakura period, Chinese styles of the Sung dynasty also influenced Kamakura wood carving. In painting as well as sculpture, Buddhist themes began to give way to more secular works; especially popular were picture scrolls (*emakimono*).

Decline of Kamakura society. During the troubled state of society at the end of the Kamakura period, the *gokenin* faced difficult times. They had borne virtually all the expense of military service against the Mongols, but their claims for reward went largely unanswered, since no lands or other wealth were confiscated from the invaders. Thus, they were financially pressed and often in debt. At the same time, important structural changes occurred in warrior houses. First, the practice of dividing lands among heirs gave way to single inheritance, often entirely to the eldest son. Power thus became concentrated in the head of the house, to whom other family members were of necessity subordinated. Second, deputies sent out by the heads of eastern warrior families to oversee their distant landholdings often broke with the main family. They formed strong ties with other local warrior houses, perhaps even becoming vassals of a *shugo*. Minimally, their ties to the Kamakura regime weakened.

General economic conditions began to undermine the position of the *bakufu* vassals. Yet, despite the social crises among the landholders, trade was flourishing. But landowners were often unable to meet their expenditures from the income of their limited holdings, even if they practiced single inheritance. Therefore, they borrowed money at high rates of interest from rich moneylenders, and many were forced to surrender their holdings when unable to repay their loans. The *bakufu* responded with debt-cancellation edicts, which gave temporary relief but neglected the long-term problem. Consequently, the gap between rich and poor became marked among the *bakufu*.

Weakening of the *bakufu*

Thus, the vassalage structure of the Kamakura regime began to unravel, and powerful local magnates, nominally Kamakura vassals, began to challenge the authority of the Hōjō regents in the *bakufu*.

The Ashikaga, Sasaki, Shōni, and Shimazu families were among the most powerful among these. Buffeted by economic changes beyond its control, the *bakufu* began to totter, shaken also by the disputes between the Hōjō family and the rival *shugo*. Subsequently, the main Hōjō house turned increasingly inward and autocratic, further alienating other vassal houses. When the Andō family raised a revolt in Mutsu province at the end of the Kamakura period, the *bakufu* found it difficult to suppress.

In addition, regional unions of small landlords developed in the Kinai (the five home provinces centered around Kyōto). Elsewhere as well, local warriors with grievances increasingly took the law into their own hands, seizing crops or otherwise disturbing local order. Termed *akutō* by the authorities, they included many different elements: frustrated local warriors, pirates, aggrieved peasants, and ordinary robbers. Cultivators as well took advantage of unsettled times to rise up against *jitō* or *shōen* proprietors.

These accumulating weaknesses of the *bakufu* prompted a movement among the Kyōto nobility to regain political power from the military. The occasion was provided by the question of the imperial succession. In the mid-13th century two competing lines for the succession emerged—the senior line centred on the Jimyō Temple in Kyōto and the junior line centred on the Daikaku Temple on the western edge of the city. In the last half of the century, each side sought to win the support of the *bakufu*. Finally, in 1318 Prince Takaharu of the junior line acceded to the throne as the emperor Go-Daigo.

THE MUROMACHI (OR ASHIKAGA) PERIOD (1338–1573)

The Kemmu Restoration and the dual dynasties. On the accession of Go-Daigo, the retired emperor Go-Uda broke the long-established custom and dissolved the office of retired emperor (*in no chō*). As a result, the entire authority of the imperial government was concentrated in the hands of a single emperor, Go-Daigo. A party of young reforming court nobles gathered around the emperor, who strove to renovate the government. But to realize his ideal of a true imperial restoration, it was necessary for Go-Daigo to rid himself of the interference of the *bakufu*. His plans for its overthrow were discovered, however, and he was arrested and exiled to Oki Island. But in the Kinai area, local leaders, supported by militant Buddhist monks, raised an army to overthrow the *bakufu*. The imperial forces were led by Prince Morinaga (or Moriyoshi) and Kusunoki Masashige, but the decisive victory was brought about by the two powerful Kantō warrior families of Ashikaga Takauji and Nitta Yoshisada, discontented vassals of the Hōjō family. In 1333 Takauji turned on the Hōjō and attacked the Hōjō headquarters in Kyōto. Yoshisada meanwhile destroyed the *bakufu* in Kamakura, at which time most of the Hōjō leaders perished in battle or by their own hand. Thus, after 140 years' rule, the *bakufu* government was brought to an end.

Overthrow of the Kamakura *bakufu*

The return of Go-Daigo to Kyōto in 1333 is known as the Kemmu Restoration. The emperor immediately set about to restore direct imperial rule. He abolished the powerful office of *kampaku* and set up a central bureaucracy. He placed Morinaga in charge of his military forces and set up members of the imperial family as provincial leaders in the north and east.

Many local warriors, however, who had joined the imperial forces in the overthrow of the *bakufu* were disappointed in the division of the spoils and the direction of the emperor's reforms. Ashikaga Takauji now turned against Go-Daigo, raising a revolt that in 1336 drove the emperor from Kyōto. Takauji enthroned an emperor from the senior imperial line, while Go-Daigo and his followers set up a rival court in the Yoshino Mountains near Nara. For the next 60 years political power was divided between the Southern Court in Yoshino and the Northern Court in Kyōto. It remained for Takauji's grandson Yoshimitsu to establish peace (1392) between the two courts; thereafter, imperial succession remained with the descendants of the Northern Court.

The establishment of the Muromachi *bakufu*. After the withdrawal of Go-Daigo to Yoshino, Ashikaga Takauji set up a *bakufu* at Nijō Takakura in Kyōto. But in 1378 Takauji's grandson, the shogun Yoshimitsu, moved the *bakufu* to the Muromachi district in Kyōto, where it remained and took final shape. Yoshimitsu, assisted by the successive shogunal deputies (*kanrei*) Hosokawa Yoriyuki and Shiba Yoshimasa, gradually overcame the power of the great military governors (*shugo*) who had been so important in the founding of the new regime. He destroyed the Yamana family in 1391, and, in uniting the Northern and Southern courts, attacked and destroyed the great *shugo* Ōuchi Yoshihiro, thus gaining control of the Inland Sea. Yoshimitsu was now raised to the highest office of prime minister, or *dajō-daijin*. He constructed the famed

Golden Pavilion (Kinkaku-ji) northeast of the capital and also reestablished trade and diplomacy with China.

Muromachi government structure. The Muromachi *bakufu* inherited almost unchanged the structure of its Kamakura predecessor. But after the appointment of Hosokawa Yoriyuki as *kanrei,* this post became the most important in the *bakufu* government. Leading officials called *shoshi* who held the additional post of *shugo* of Yamashiro province (now in Kyōto urban prefecture) were next in importance to the *kanrei.* In local administration, a special administrator was set up in Kamakura to control the 10 provinces of the Kantō area. The 11 provinces of Kyushu were placed under control of an office known as the Kyushu *tandai.*

Changing role of the *shugo*

The crucial difference between the two *bakufu,* however, was the difference in the role of the *shugo.* Appointed first by Takauji, many rose to positions of great power. By Yoshimitsu's time, the structure of the *bakufu* was essentially a delicate balance between the Ashikaga shogunal house and about a dozen major *shugo* houses. Yoshimitsu made them all establish primary residence in Kyōto, and they often sent out deputies to manage their provincial areas in their absence. Consequently, in later years many powerful *shugo* were overthrown by their own deputies.

In the Kamakura period the authority of the *shugo* was essentially limited to security matters. In the latter half of the Northern and Southern courts period, their executive power over the areas under their control was increased. As the number of disturbances grew, they gained wide powers of military command. Many *shugo* succeeded to their domains by inheritance, and in cases such as that of the Yamana family a single *shugo* sometimes held a number of provinces. If the primary agent of the Kamakura *bakufu* had been the *jitō,* the *shugo* was the defining office of the Muromachi regime. From the outset, the controlling power of the Ashikaga *bakufu* was relatively weak, and, especially after the death of Yoshimitsu, the tendency for powerful *shugo* to defect became marked. Hence, as time passed the office of shogun became increasingly impotent.

The growth of local autonomy. In the villages around Kyōto, the status of farmers rose markedly as agriculture became more highly developed, and commerce and small-scale manufacturing prospered. Also, confederations of the middle and small landlords, or *myōshu,* proceeded apace and often led to uprisings against absentee control. In such confederations, a leader called the elder (*otona*) would be selected to head village government.

As self-government became strong in the communities, the resistance of farmers became fierce. After the unification of the Northern and Southern courts, armed uprisings broke out among the farming villages, the peasants demanding reductions in yearly taxes from the old proprietors and a moratorium on debts owed to the money-lenders. A large-scale uprising of this kind took place in 1428 in the last years of Yoshimitsu's rule. In 1441 farmers living around Kyōto attacked the pawnbrokers and demanded that the *bakufu* declare a moratorium on debts. Thereafter, uprisings occurred on a greater or lesser scale almost yearly.

Trade between China and Japan. Trade with Ming dynasty China began after the *bakufu* agreed to suppress Japanese piracy. Ashikaga Takauji had sent ships of the Tenryū Temple to trade with the Yüan (Mongol) dynasty. But trade then ceased because of the internal disturbances, and pirates from the maritime districts of western Japan raided both China and the Korean peninsula. Yoshimitsu began formal trade relations with Ming China and Korea, repatriating a large number of Chinese who had been taken captive by the pirates. In response, the Ming also began to trade with Japan. In order to distinguish between pirate ships and trading ships, seals received from the Ming called *kangōfu* were used, hence the use of the term *kangō,* or tally, trade.

The *kangō* trade

Profits from the China trade were important to the *bakufu,* but control of this trade later came into the hands of the western *shugo* families of the Hosokawa and Ōuchi. After the Ōnin War (see below), the Ōuchi controlled the trade, but with the destruction of the Ōuchi the *kangō* trade ceased and piracy again became rife. Trade with Chosŏn (Yi) dynasty Korea was carried on through the agency of the Sō family of Tsushima. Japanese traders even established settlements in southeastern Korea, including Pusan. Also included in the trade with China and Korea were goods imported by Japanese merchants from the Ryukyu Islands and from the South Seas.

The Ōnin War (1467–77). During the rule of the shogun Ashikaga Yoshimasa a general civil war broke out in the area around Kyōto, caused by economic distress and precipitated by a dispute over the shogunal succession. Lacking children of his own, Yoshimasa at first proposed that his younger brother should succeed him. But when he later fathered a child a serious dispute arose over control of the Ashikaga family. The two chief administrators, Shiba and Hatakeyama, and most of the remaining *shugo* also took sides in the power dispute, with Hosokawa Katsumoto and Yamana Sōzen (Yamana Mochitoyo) at the head. In 1467, the first year of the Ōnin era, fighting broke out between the "eastern" army of the Hosokawa party and the "western" army of the Yamana faction. The eastern army had the advantage of the support of both the emperor and the shogun, but the western army, assisted by the Ōuchi family, recovered its power, and fighting raged mainly in and around Kyōto. After 11 years the war itself ended, but the fighting spread to the provinces. As a result, farming villages frequently mounted armed uprisings in self-defense. The leaders of these uprisings were local samurai with village roots. Such men frequently established themselves as domain lords (daimyo) during the disturbances. They formed associations and often mounted uprisings that extended over an entire province and challenged the great *shugo.*

During this constant warfare, the civil aristocracy and temple complexes lost much of their income from *shōen,* which, in any event, had been declining. Many of them left the capital. This migration of aristocrats and priests functioned to diffuse the higher culture of the capital to the provinces. Old traditions were destroyed, but from the ashes a new culture was born. The shogun Ashikaga Yoshimasa ultimately turned his back on a troubled world and built a detached residence—the Silver Pavilion (Ginkaku-ji)—in the Higashiyama section of Kyōto, paying little attention to matters of government. The political power of the *bakufu* thus became virtually nonexistent, and real power came into the hands of the chief administrators of the Hosokawa family (1490–1558). In the 16th century actual power devolved into the hands of their retainers, the Miyoshi family (1558–65), until it was finally usurped by their own retainers, the Matsunaga family (1565–68).

The Sengoku ("Warring States") period. *The emergence of new forces.* After the Ōnin War, the power of independent local leaders increased markedly, and in many instances deputies of great *shugo* houses usurped the domains of their superiors, retainers overthrew their overlords, and branch families seized power from main families. Because of this tendency, the previous *shugo* almost completely disappeared from Kyōto and the surrounding provinces; a new type of domain lord, the daimyo, took their place. Since this time was marked by constant warfare among many such lords, it is called the Sengoku ("Warring States") period, named for a somewhat similar period in ancient Chinese history.

Rise of the daimyo

Until the first half of the 16th century, daimyo in the various localities were thus building up strong military bases. During this period, the provinces held by the daimyo were almost completely free of *bakufu* control. The daimyo turned local leaders into their retainers, taking away their independence by enforcing land surveys and directly controlling the farming villages. Daimyo such as the Imagawa, Date, and Ōuchi issued their own laws, called *bunkokuhō,* to administer their own territories. In farming villages the daimyo, in addition to carrying out detailed land surveys, also built irrigation dikes and opened new rice fields in order to stimulate production. To concentrate their power they also readjusted the disposition of local fortified strongholds, gathered their retainers into castles, and reorganized roads and post stations to centre on their castle towns (*jōkamachi*).

Commerce and towns made marked development at this time in Japan's history. Periodic markets also sprang up throughout the country. In large cities such as Kyōto, commodity exchange markets were set up to handle huge quantities of rice, salt, fish, and other goods; wholesalers, or *toiya,* specialized in dealings with distant areas. The circulation of coined money also became vigorous. Muromachi guilds showed a strong monopolistic tendency in trying to protect themselves against new-style merchants who emerged, while new guilds were set up in the castle towns under the direct control of the daimyo.

Among the cities of the time, next to Kyōto and Nara, Uji-Yamada, Sakamoto, and other towns sprang up outside the gates of major temples and shrines. Besides these, towns naturally grew up around the castles of the daimyo. As the castles shifted from serving as defensive mountain fortresses to administrative strongholds in the plains, markets were opened outside the castle walls, and merchants and artisans gathered there to live. Harbour towns (*minato machi*) such as Sakai, Hyōgo, and Onomichi on the Inland Sea, Suruga and Obama on the Sea of Japan, and Kuwana and Ōminato on Ise Bay also flourished as exchange centres. Sake brewers, brokers, and wholesale merchants were leading townsmen (*machishu*), and town elders (*otona*) were chosen to carry on local government through assemblies.

The arrival of the Europeans. As the warring daimyo carved out their territories, neither emperor nor shogun was able to govern the domestic scene, let alone control overseas trade. It was at this point in Japanese history that the Spanish and Portuguese made their appearance in the archipelago. In 1543 several shipwrecked Portuguese were the first Europeans to arrive in Japan, and the art of musket construction they passed on at this time immediately spread. This new technology, eagerly sought by the daimyo, revolutionized warfare in Japan. In 1549 the Jesuit missionary Francis Xavier arrived in Kagoshima. After missionary work for more than two years, he left Japan; but thereafter Jesuit missionaries arrived continuously. The missionaries utilized trade in goods from the Portuguese ships to propagate Christianity. Thus, the daimyo of the Sengoku era, seeking profits of foreign trade and the acquisition of military equipment and supplies, protected Christianity. Some daimyo became Christian converts. Farmers also increasingly became converts.

The Jesuit missionaries

The establishment of warrior culture. While absorbing the traditional culture of the civil aristocracy, the warrior houses that established themselves in Kyōto during Muromachi times also introduced the continental culture of the Sung, Yüan, and Ming dynasties, especially the culture associated with Zen Buddhism, thus fashioning a new warrior culture. This process began with the golden age of Ashikaga Yoshimitsu at the end of the 14th century. Renga (linked verse) and nō drama flourished. The essence of this culture found concrete expression in Yoshimitsu's Golden Pavilion at Kitayama ("Northern Mountain"). Kitayama culture, while absorbing new Zen influences from China, retained much of the earlier native aristocratic culture.

The era of the shogun Yoshimasa, following the destruction caused by the Ōnin War, was one of an even deeper Zen flavour and showed a refined appreciation of simplicity and quiet profundity. Yoshimasa's somber Silver Pavilion and its garden in eastern Kyōto truly reflect the essence of the polished style of Higashiyama ("Eastern Mountain") culture. Higashiyama culture was the forerunner of the Azuchi-Momoyama and Edo cultures.

In Buddhism, the great ancient temples like the Enryaku Temple became mere shadows of their former greatness with the gradual diminution of their *shōen.* Since the Kamakura period, the new Rinzai Zen sect had been especially favoured by high-ranking warrior houses. The Muromachi shogunal family (the Ashikaga) gave special protection to followers of the priest Musō Soseki of this sect, which flourished in the Gozan monasteries (the five most important Zen monasteries) in Kyōto. But the Gozan monasteries ceased to prosper as the *bakufu* declined. In contrast, the Myōshin and Daitoku temples—also of the Rinzai sect but outside the Gozan system—rose to prominence, the latter perhaps best known for the work of the monk Ikkyū.

It was during this period that Rennyo (1415–99) of the Shin (True) sect of Pure Land Buddhism rose to prominence, and the Hokke (Lotus) sect gained adherents among warriors and merchants. Moreover, it was during this time that the custom of pilgrimages to the holy places of the Buddhist deity Kannon, to the Shintō shrines at Ise, and to the summit of Mount Fuji also became popular. Accompanying this trend was the development of a worldly Shintō belief.

Pilgrimages

In the arts the nō drama developed in the Kamakura period out of the older tradition of agricultural festival dances, and guilds (*za*) were formed to serve at the ceremonies of temples and shrines and at funeral services. *Kyōgen* (dialogue plays with dance), which developed from the comic elements of an older form of entertainment called *sarugaku,* were performed in the intervals of nō drama. Traditional Japanese waka verse was still composed, but renga (linked verse) became ever more popular and was enjoyed by the warriors and the common people alike. After a time, however, even renga became overly formal, and the free-style verse called haikai was born.

As Zen prospered, the *shoin* architectural style closely connected with this school was widely adopted by both warriors and civil aristocrats in the construction of their residences, becoming the foundation of present-day Japanese domestic architecture. In the construction of gardens, delight was taken in adding the Zen mood of retreat from the world to the old *shinden* style, making symbolic use of streams, flowers, and bushes. Later, even more symbolic gardens were constructed using arrangements only of stones, raked sand, and gravel.

The carving of images of the Buddha and the Buddhist paintings that had flourished in the Kamakura period declined in later Muromachi times. Yamato-e painting also declined, and the picture-scrolls lost their freshness. In their place, the increased interest in Zen led to the introduction of monochrome painting in the Sung and Yüan style. The father and son Kanō Masanobu and Kanō Motonobu later introduced the gentle forms of Yamato-e to monochrome painting and became the founders of the new Kanō school.

Tea drinking, introduced from Sung China in the Kamakura period, spread among warriors and even common people from the mid-14th century. In the time of the shogun Yoshimasa, Murata Shukō, a man of merchant background from Nara, began the *wabi-cha* form of tea ceremony. This new form spread among the warriors and great merchants and was further stylized by the Sakai merchant Takeno Jōō. The development of the tea ceremony stimulated new forms in tearoom architecture, flower arrangement, pottery, and even the Japanese cakes served with tea. The Higashiyama cultural tradition was further diffused among the common people, and as the levels of wealth and education of urban merchants and artisans rose, they, too, came to enjoy nō and *kyōgen* dramas, the tea ceremony, and renga. Muromachi fiction celebrated the life of the burgeoning artisan and merchant classes. Thus, while warfare was rife in the Muromachi period, it gave Japan some of its most distinctive cultural institutions. (T.T./G.C.H.III)

The tea ceremony

Early modern Japan (1550–1850)

UNIFICATION

The Oda regime. In the 1550–60 period the Sengoku daimyo, who had survived the wars of the previous 100 years, moved into an even fiercer stage of mutual conflict. These powerful daimyo were harassed not only by each other but also by the rise of common people within their domains. The daimyo sought to resolve their dilemma by acquiring land and people to widen their domains and, finally, by trying to seize control of the whole country. That, of course, required the control of Kyōto, the political centre of Japan since ancient times. Out of these bloody struggles emerged one Sengoku daimyo, Oda Nobunaga of Owari province (in modern Aichi prefecture), who succeeded in occupying the capital as the first feudal unifier.

The emergence of Nobunaga's regime reversed the feudal disintegration of the previous century and moved the country toward unification. Oda was a military genius, who was the first to successfully adapt firearms to Japanese warfare. His bold wars of suppression, waged against both other daimyo and recalcitrant religious communities, led to a great redrawing of the political map of Japan, previously split up among daimyo throughout the country. In the Kinai district, where Nobunaga's conquered territory was centred, however, he established control by dividing his new domain among his commanders. Rather than completely abrogating the long-established privileges of the temples, shrines, and local landlords (*kokujin*), he at first recognized them. Cadastral surveys aimed at strengthening feudal landownership were at this stage carried out to define the size of fiefs (*chigyō*) of Nobunaga's retainers in order to confirm the extent of their obligations to him.

Nobunaga's unification policy was predicated on a separation of warriors from the farmers, but unification was hampered because of resistance from old political forces, especially several major Buddhist temples. Unification proceeded further during the era of Nobunaga's successor, Toyotomi Hideyoshi.

The Hideyoshi regime. Nobunaga's father was a minor Owari daimyo, whereas Hideyoshi was the son of a peasant from the same province. After entering Nobunaga's service, Hideyoshi impressed all with his brilliant talents, and he soon rose to become one of Nobunaga's most powerful commanders. After Nobunaga's death, Hideyoshi eliminated many rivals by relying on his superb political judgment and shrewd actions, firmly establishing himself as successor. Following in Nobunaga's footsteps, Hideyoshi proceeded to unify the whole country at a rapid pace, and by 1590 all Japan—from Kyushu in the southwest to Tōhoku in the northeast—had come under his control.

At the core of Hideyoshi's unification policy was its firm establishment in the principle of the separation between warriors and peasants. Hideyoshi adopted several major policies to accomplish this end: a comprehensive land survey (*kenchi*), the disarmament of the peasantry, and the separation of the classes. The so-called Taikō land survey was carried out throughout the country from 1583 to 1598, being completed just before Hideyoshi's death. As a result of this survey, the complicated relationships of rights to landownership that had developed since the Kamakura period were now clarified. Landowning relations were now based on *kokudaka—i.e.,* on the actual product of the land. The *kokudaka* system also applied to the landholdings of the daimyo for distribution among their retainers. In place of previous land taxes (*nengu*), an assessment of *kokudaka* was made as so many hundred or ten thousand *koku* of rice. A *koku* represented the amount of rice consumed by one person in one year (about five bushels); the amount also was used as a standard on which military services were levied in proportion.

The Taikō land survey

As part of the process, a register was drawn up in every village. Peasants had their rights as cultivators recognized to the extent that their land was duly registered; in return, they were bound to pay land taxes in rice and were forbidden to neglect the cultivation of their fields or to move elsewhere. The promulgation of an order of social-status control in 1591 prohibited warriors from taking up farming and forbade other daimyo from employing a samurai who left his master. The ordinance also forbade artisans and merchants from residing in villages, thus extending Nobunaga's attempt to separate warriors and farmers into a social-class system of warriors, farmers, artisans, and merchants. Hideyoshi's so-called "sword hunt" (*katana-gari*) of 1588, which attempted to disarm the peasantry, was an important prerequisite for this policy. With the establishment of the *kokudaka* system, the Taikō land survey delivered the final blow to the *shōen* system of manorial holdings, which had already virtually disappeared under the onslaught of the Sengoku daimyo. The feudal *chigyō* system, based on the *kokudaka* assessment, was established throughout the country. The provincial daimyo all submitted to Hideyoshi's regime, and the more egalitarian, alliance-like relationship between Nobunaga and the former Sengoku daimyo was replaced by a clear lord-vassal relationship.

The political structure of the Hideyoshi regime was not yet fully sufficient, however, to be the unified governing authority for the whole country. For example, the *kurairechi* (lands under its direct control), which were the immediate financial base of the regime, amounted to more than 2.2 million *koku* by the time of Hideyoshi's death, nearly one-eighth of Japan's cultivated land. But aside from those in the metropolitan and surrounding provinces, these lands were in many cases divided among the distant, independent *tozama* ("outside") daimyo, and the management of these lands was entrusted to them. Such lands were thus not firmly in the grasp of the regime. This limitation of Hideyoshi's regime gave rise to internal power struggles and finally drove Hideyoshi to such reckless actions as the invasions of Korea in 1592 and 1597. These two ill-advised adventures were designed to bring China under Hideyoshi's sway and to provide an outlet for tens of thousands of warlike samurai. Hideyoshi's regime collapsed on the failure of the second Korean expedition and as the direct result of Hideyoshi's subsequent death. Hideyoshi failed to bequeath his power to his heir, Hideyori, and Tokugawa Ieyasu emerged as the strongest candidate to succeed him.

Invasions of Korea

Azuchi-Momoyama culture. Cultural historians often refer to the last few decades of this era as the Azuchi-Momoyama period, taking the name from Oda Nobunaga's massive fortress at Azuchi, overlooking Lake Biwa at Hikone, and Hideyoshi's magnificent edifice in the Momoyama district, southeast of Kyōto. Often abbreviated as, simply, the Momoyama period, it is characterized by gaudy splendour celebrating the ego of the two great rulers. The defining feature of the age is the castles that dotted the landscape between the 1580s and 1630s. Many of the associated castle towns were the forerunners of Japan's present provincial capitals.

The castles were often filled with items reflecting the personalities of the rulers. In particular, Momoyama culture is noted for the magnificent standing screens, *fusuma* (sliding doors), and wall paintings of a monumental nature that decorated the castles. Artists of the Kanō school produced colourful pictures of animals and landscapes. Momoyama paintings provide a vivid contrast to the somber tones of the monochrome paintings of the Muromachi era. A specific genre within this tradition is often referred to as *namban* ("southern barbarian") pictures, since they represent both the European priests and traders—referred to as "southern barbarians" since they had entered Japan from the South Seas—of the day and their magnificent ships. Nobunaga and Hideyoshi spent great amounts of time and money indulging their cultural proclivities, especially the tea ceremony (*chanoyu*). Hideyoshi extended his patronage to the tea master Sen no Rikyū, the figure from whom all current tea masters trace their lineage.

THE BAKUHAN SYSTEM

The establishment of the system. The ancestors of Tokugawa Ieyasu, the founder of the Edo *bakufu,* were the Matsudaira, a Sengoku daimyo family from the mountainous region of Mikawa province (in present Aichi prefecture). But when they were attacked and defeated by the powerful Oda family from the west, Ieyasu's father, Hirotada, was killed. After his father's death Ieyasu was sent to the Imagawa family and spent 12 years there under detention. When, in 1560, Oda Nobunaga destroyed the Imagawa family in the Battle of Okehazama, Ieyasu was finally released. Ieyasu returned to Okazaki in Mikawa and brought this province under his control. By 1582 he was a powerful daimyo.

Rise of Tokugawa Ieyasu

When Hideyoshi seized power, Ieyasu at first opposed him. But he then submitted, and, rising to be the most powerful daimyo among Hideyoshi's vassals, he became chief of the five *tairō* (senior ministers), the highest officers of the Hideyoshi regime. After Hideyoshi's death the daimyo split between those supporting Hideyori and those siding with Ieyasu. Matters came to a head at the Battle of Sekigahara in 1600, where Ieyasu won a decisive victory and established his national supremacy. Ieyasu had seen the failure of both Nobunaga and Hideyoshi to consolidate a lasting regime, and in 1603 he set up the Edo *bakufu* (more commonly known as the Tokugawa shogunate [1603–1867]) to legalize this position. Assuming the title shogun, he exercised firm control over the remaining daimyo at this time. Furthermore, Hideyoshi's son and heir Hideyori was reduced to the position of a daimyo of the Kinki (Ōsaka area) district. Two years after the establishment of the *bakufu,* Ieyasu relinquished the post of shogun to his son Hidetada, retiring to Sumpu (modern city of Shizuoka) to devote himself to strengthening the foundations of the *bakufu.* In 1615 Ieyasu stormed and captured Ōsaka Castle, destroying Hideyori and the Toyotomi family. Immediately afterward, the Laws for the Military Houses (Buke Shohatto) and the Laws for the Imperial and Court Officials (Kinchū Narabi ni Kuge Shohatto) were promulgated as the legal basis for *bakufu* control of the daimyo and the imperial court. In 1616 Ieyasu died, the succession already having been established.

Under the second and third shoguns, Hidetada and his successor, Iemitsu, the *bakufu* control policy advanced further until the *bakuhan* system—the government system of the Tokugawa shogunate; literally a combination of *bakufu* and *han* (the domain of a daimyo)—reached its completion. Confiscations and reductions of domains and wide-scale transfers of daimyo also took place, distributing the strategic districts of Kantō, Kinki, and Tōkaidō among the daimyo who were relatives and retainers of the *bakufu,* thus keeping the "outside" (*tozama*) lords in check. Along with the rearrangement of the daimyo, the lands under the direct control of the *bakufu* also were increased at key points throughout the country. The most important cities—Kyōto, Ōsaka, and Nagasaki—and mines (notably, the island of Sado) also were placed under direct *bakufu* administration.

The *sankin kōtai* system

The *bakufu* also revised the Laws for the Military Houses and established a system called *sankin kōtai* (alternative attendance), by which the daimyo were required to pay ceremonial visits to Edo every other year, while their wives and children resided permanently in Edo as hostages. The system also forced the daimyo—especially the potentially dangerous *tozama* who lived farthest away—to spend large sums of money to support two separate administrative structures and trips to and from Edo. Tokugawa *bakufu* domains now amounted to more than seven million *koku*—about one-fourth of the whole country. Of these lands, more than four million *koku* were under its direct control, and three million *koku* were distributed among the *hatamoto* and *gokenin,* the liege vassals to the *bakufu.* In addition, because the *bakufu* declared a monopoly over foreign trade and alone had the right to issue currency, it had considerably greater financial resources than did the daimyo. In military strength as well, it was also far more powerful than any individual daimyo.

In step with the structural organization of the *bakufu* as the supreme power, the domain administration (*hansei*) of the daimyo also progressively took shape. The relationship between the shogun and the daimyo was that of lord and vassal, based on the feudal *chigyō* system. In theory, the land belonged to the shogun, who divided this among the lords as a special favour, or *go-on.* In return, the daimyo incurred the obligation to provide military and other services to the shogun. Precisely the same connection existed between the domain lords and their retainers. In order to restrict the traditional right of their vassals to *chigyō,* or subdomains, daimyo rewarded them instead with rice stipends (*kuramai*), thus increasing their dependence on the daimyo. At the same time, this policy increased the lands under the direct control of the daimyo, strengthening the economic base of the domain. Thus, the daimyo employed the same methods toward their own vassals as the *bakufu* used to control them. In this way, a hierarchical, "feudal" regime was established by means of the *kokudaka* system.

Control over the agricultural populace was now further strengthened. The Taikō land survey had recognized the rights of the peasants as actual cultivators of the land and made them responsible for taxes. Similar in intent, the land surveys of the *bakufu* and the daimyo were much more detailed and precise, concerned, as they were, with extracting the greatest possible tax yield. Tokugawa villages were composed of a main core of small farmers, generally called *hyakushō.* The inhabitants of towns and villages throughout the country were required to form *gonin-gumi* ("five-household groups"), or neighbourhood associations, to foster joint responsibility for tax payment, to prevent offenses against the laws of their overlords, to provide one another with mutual assistance, and to keep a general watch on one another. Economic controls over peasants were further strengthened. They were strictly prohibited from buying, selling, or abandoning their land or from changing their occupation.

Control of the villages

The enforcement of national seclusion. The 1630s also marked an important dividing line in foreign relations with the issuance of a series of directives enforcing a policy of national seclusion, later called *sakoku* (literally, "closed country"). The seeds of this policy had been sown in trade control and in measures against Christianity by the Nobunaga and Hideyoshi regimes. Hideyoshi, although strongly attracted to trade as a source of national wealth and military strength, had issued an order for the exclusion of the missionaries. Ieyasu, even more strongly attracted by profits, made efforts to trade not only with the Portuguese Roman Catholics but also with Protestant Holland and England, protecting trade with the southern regions by granting special licenses, or *shuin-jō* ("red-seal license"), to oceangoing merchant ships. But Ieyasu's encouragement of trade was aimed at establishing a *bakufu* trade monopoly.

Eager for trade, Ieyasu was initially tolerant of Christian proselytization, but later he came to fear that the Christians would join Hideyoshi's heir Hideyori to resist the *bakufu,* and he took steps to prohibit Christianity before his destruction of the Toyotomi family. Decrees prohibiting Christianity were promulgated in 1612 and 1614, and the persecution of its adherents began immediately thereafter. Persecution became much more severe under Hidetada and Iemitsu, until, at length, it became official policy to stamp out Christianity even at the sacrifice of trade. This policy became manifest with the seclusion orders of the 1630s.

The Shimabara Rebellion

In 1637, in resistance to heavy taxes and the prohibition of Christianity, Amakusa Shiro, a Christian masterless samurai (*rōnin*), led an uprising of peasants and Christians in the Shimabara Peninsula of Kyushu. For five months they put up a fierce fight before their defeat by the *bakufu* army. The system of registration at Buddhist temples was instituted: all Japanese were required to register as parishioners to a parent Buddhist temple, which every year had to guarantee that the parishioner was not a Christian. When in 1639 Portuguese ships were forbidden to visit Japan, the *sakoku* orders were completed. The Dutch and the Chinese were allowed to trade as before, although this trade was restricted and confined to the island of Dejima at Nagasaki. Iemitsu also allowed a certain amount of trade with Korea and the Ryukyu Islands.

Scholars continue to debate the effects of national seclusion, but its impact on Japan was profound. The vigorous desire of the Japanese of the Sengoku era to expand overseas was thenceforth transformed into an attitude hostile to foreign trade, if not to foreigners themselves. On the one hand, the seclusion policy was instrumental in enabling the Tokugawa *bakufu* to establish a prolonged peace of nearly 300 years; yet on the other, it has been argued that this simply prolonged a rigid feudal system to an extent unknown elsewhere in the world.

The Tokugawa status system. Thus, the *bakuhan* system was firmly solidified by the second half of the 17th century. The establishment of a strict class structure of warriors, farmers, artisans, and merchants (*shi-nō-kō-shō*) represents the final consummation of the system. Distinctions between the statuses of warriors, farmers, artisans, and merchants were strictly enforced, but the distinction between the samurai and the other three classes was especially strict. Forming barely 7 percent of Japan's total population, warriors levied taxes on the farmers, who formed more than four-fifths of the population and who thus provided the economic foundation of the system. Symbolizing their dominance of society by force of arms, samurai wore two swords; by law, the other classes were forbidden to wear them. Concern for strict status differentiation was evident even in family relationships, as absolute obedience was demanded from members of the family toward the house head (*kachō*). Among the family members, the status of women was especially low.

The establishment of the Tokugawa regime created a need for legitimation, a new worldview, and a system of ethics to support it. Neither the Shintō nor the Buddhist ideologies of the earlier medieval society was adequate. But the ideas of Neo-Confucianism, especially of the Sung dynasty Chu Hsi school (Shushigaku)—which had been well-known to political and ethical thinkers since the 13th century—provided an intellectual rationalization for the status-oriented social structure of the *bakuhan* system. Shushigaku appealed especially to the feudal rulers because, among the various schools of Confucianism, it was the most systematic doctrine. The Chu Hsi scholar Hayashi Razan served as advisor to the first three shoguns. He is said to have had a hand in the drafting of all *bakufu* official documents and in the formulation of *bakufu* laws. His political ideas provided a historical justification for the establishment of the Tokugawa shogunate, based upon the concept of *tendō* ("way of heaven"). *Tendō* essentially took on the connotation of the Chinese term *t'ien-ming* ("mandate of heaven"; Japanese: *tenmei*), and Razan and other Confucian thinkers provided an explanation and justification for changes in rulers through the process of *gekokujō* (overthrow of superiors by inferiors) of the Sengoku period. But the role of Chu Hsi political-ethical thought in Tokugawa times was to repudiate the revolutionary idea of *gekokujō* by stressing the legitimacy of Ieyasu's new regime, emphasizing instead the idea of *kenshin* ("devotion," or "loyalty"), linking this to Confucian moral concepts. Razan argued that the separation of the four classes of society was in accord with the teachings of Confucius. The two central moral ideals of Confucianism were *chū*, or "loyalty," and *kō*, or "filial piety." Tokugawa thinkers like Razan placed more emphasis on *chū* as a support for feudal lord-vassal relations than on *kō*, which was a family ethic. Orthodox Chu Hsi thought was a perfect conservative philosophy of statecraft that valued loyalty and order above all else.

Hayashi Razan

Japanese thinkers of the 17th century could hardly have been expected to fully ingest a foreign political philosophy already several hundred years old, and challenges to orthodox Chu Hsi thought were many. Some argued for a return to the original teaching of Confucius himself. The philosophy of another Sung thinker, Wang Yang-ming, also held a special place in Confucian circles in the early Edo period. Wang Yang-ming studies (Ōyōmeigaku in Japanese) were characterized by a strong subjective idealism but, at the same time, were quite practical since they emphasized the unity of thought and deed. Virtue had to be not only cultivated in the abstract but practiced as well. Nakae Tōju, often regarded as the father of Japanese Wang Yang-ming studies, was so earnest in performing virtuous acts that he was called the sage of Ōmi. One of his followers, Kumazawa Banzan, transformed Wang Yang-ming studies from a means for individual spiritual training into a method for political reformation.

Commerce, cities, and culture. By reducing Ōsaka Castle and quelling the Shimabara Rebellion, the Tokugawa regime brought to an end the period of violence and ushered in an era of unprecedented domestic peace. As a result, commerce was promoted and cities developed. Widespread commercialization occurred in the latter half of the 17th century, centred in the Kinki region, where productive capacity was the most advanced. Now the nationwide farming populace (*hyakushō*) of independent landowners, although subject to heavy taxes and various kinds of labour services, sought the means to enjoy a better standard of living. In addition to their primary efforts as cultivators, they reclaimed new lands and produced various commercial crops and handicraft goods for sale in the city and town markets. Communications and transportation also developed for the circulation of such goods, thanks to the earlier efforts of various daimyo to maximize production in their domains and to the increased mobility caused by the *sankin kōtai* system. As a result of the development of commerce and communications, new-style merchants such as wholesalers and brokers came to the fore, and powerful financiers also appeared.

Urban growth

There was a massive growth of urban centres in the first half of the Edo period, mainly represented by the castle towns of the various daimyo. These daimyo, numbering some 250 for most of the period, were allowed by the *bakufu* to have but one castle, and thus there was a move to pull down other castles and concentrate the samurai of each *han* in a capital castle town. These castle towns gradually came to acquire the character of commercial cities, as some farmers abandoned the countryside and merchants emerged to serve the needs of the burgeoning urban population. Purely commercial cities and post towns (towns along highways) also arose throughout the country as part of this massive urbanization. The three main cities of Edo, Ōsaka, and Kyōto, under the direct control of the *bakufu*, were especially developed.

The early and mid-Edo periods produced many remarkable figures in the fine arts and crafts. Perhaps the three artists most representative of the culture were Ihara Saikaku in *ukiyo-zōshi* ("tales of the floating world") genre novels, Chikamatsu Monzaemon in *jōruri* ("puppet play") drama, and Matsuo Bashō in haiku poetry. All three flourished during the Genroku era (1688–1704), the name more broadly denoting a golden age of cultural development roughly 50 years long during the late 17th and early 18th centuries. Saikaku was an Ōsaka townsman who first aspired to write haikai—humorous renga (linked-verse) poetry from which the more serious haiku was derived—and for more than 30 years he was active as a haikai composer. By 1682 Saikaku largely had given up haikai, however, and began to write *ukiyo-zōshi,* producing about 20 masterpieces in succession, beginning with *Kōshoku ichidai otoko* (1682; *The Life of an Amorous Man*). The unique urban spirit of the age can be seen in the word *ukiyo,* which had meant "sad world" in Buddhist terms during medieval times. Written with a different Chinese ideogram in Edo times, it now came to mean "floating world" and implied pleasure—specifically from the pleasure quarters of the great Edo cities.

Matsuo Bashō became closely attached to haiku (although the word itself was not coined until the 19th century) and fashioned it into a popular form of poetry. Bashō began writing hokku (17-syllable opening verses for renga) as separate poems, developing a new style called *shōfū* or "Bashō style." He essentially grafted the aristocratic conceptions of medieval poetry onto the more mundane feelings of Tokugawa urban culture, creating a highly popular poetic form.

About the turn of the 17th century, the *Jōrurihime monogatari* (a type of romantic ballad), which drew on the traditions of the medieval narrative story, was for the first time arranged as a form of dramatic literature accompanied by puppetry and the samisen (a lutelike musical instrument). It continued to develop until the three great masters—Takemoto Gidayū as narrator, Chikamatsu Monzaemon as composer, and Tatsumatsu Hachirobei as puppeteer—made *jōruri* into a highly popular Tokugawa performing art, enjoyed by all classes of society. Chikamatsu wrote more than 80 *jidaimono* (historical dramas) and 20 *sewamono* (domestic dramas focusing on urban society), both for *jōruri.* He also wrote more than 30 kabuki plays. The chief theme running through Chikamatsu's works is the idea of *giri* ("duty"), the traditional consciousness of honour and dignity in one's motives and of social consciousness in human relations. *Giri* is constantly in conflict with *ninjō* ("human feelings," especially love), and this tension provides the drama in many of his works.

Kabuki

Besides the licensed quarters for prostitutes, theatrical districts also flourished in the Genroku era. Kabuki drama also developed in the early Edo period. *Okuni* kabuki, named for the female dancing troupe led by Izumo Okuni, became popular at the turn of the 17th century and is conventionally regarded as the origin of this dramatic form. Kabuki subsequently developed from its previous dancing-act form into a theatrical form centred on a dramatic plot with realistic acting. In western Japan (Kyōto and Ōsaka), the style that emerged was called *wagoto* ("tender business"), which had a pronounced comical element and concentrated on love; by contrast, the popular form of Edo kabuki was *aragoto* ("rough business"), which focused on the rash actions of historical heroes.

Despite the popularity of these new theatrical forms, traditional arts of nō drama, the tea ceremony, and flower arrangement also reached new stages of development in the period. The tea ceremony (*chanoyu*) in particular became popular and was practiced not only by the shogun and daimyo but also by the newly risen merchants, who used their wealth to become eager collectors of famous antique tea-ceremony utensils. As the tea ceremony became popular, many schools emerged, most notably the Sen-ke (Sen house), the school of Sen Rikyū. The art of the tea ceremony came to be monopolized by the house heads of the various schools, fostering the development of the "profession" of tea master.

Distinctive development also occurred in the fine arts and crafts. Ogata Kōrin, for example, brought decorative painting to its highest stage of perfection, bequeathing to posterity many splendid masterpieces in gold lacquer (*maki-e*) and other media. Techniques of dyeing and weaving were also improved, and the weaving and decorating of the traditional kimono became more colourful. In Edo, Hishikawa Moronobu developed the Chinese technique of wood-block printing to produce the ukiyo-e ("pictures of the floating world") style, which met a growing popular demand. Many great Edo-period artists—*e.g.,* Andō Hiroshige and Katsushika Hokusai—developed the ukiyo-e genre into a unique Japanese art form. Famous centres of pottery production also flourished at various places throughout the country, some ancient, like Seto, but others, like Hagi, stimulated by the influence of Korean potters captured during Hideyoshi's invasions.

Pottery production

Both the old ceremonies of the imperial court and the various forms of warrior etiquette developed by the successive *bakufu* were codified, studied, and even extended to the common people, helping to shape manners throughout the country. Indeed, Japanese customs in dress, food, and housing became established and somewhat standardized during the Edo period. Even eating habits changed from two to three meals a day; in the cities rice became the standard food, and a rich variety of cakes and sweets were consumed by urban dwellers.

THE WEAKENING OF THE BAKUHAN SYSTEM

As Japan entered the 18th century, the *bakuhan* system began to show signs of weakness. The finances of both the *bakufu* and the *han* were theoretically based on a rice-producing economy, in which administrators endeavoured to levy taxes to be paid in kind, mostly in rice, centred on the annual crop. Rice and other crops were then transported to the great central cities of Edo and Ōsaka, where they were exchanged for money. The extremely diverse economic and social life of these cities was based upon a money economy in which people and produce were constantly exchanged. This activity radiated outward to the various daimyo castle towns and, inevitably, into the countryside as well. Thus, even the rural areas of Japan were increasingly drawn into a monetized economy, and peasants everywhere paid part of their taxes in money. If commercial development had been largely a phenomenon of the cities in the 17th century, in the 18th and 19th centuries it spread to the hinterlands of Japan, where small-scale producers of goods, distributors, and even retailers appeared. Inevitably, it meant the rise of some wealthy members of the rural populace, who used their wealth to invest in land and commercial ventures and to "ape their betters" in the cities in both custom and culture. Few farmers, however, prospered through producing commercial goods, and the majority of peasants remained impoverished. Rural villages were characterized by a few wealthy farmers, a majority of small-scale independent landholders, and a growing number of impoverished tenants. Many small-scale farmers, squeezed by the demands of commercial development, were forced to part with their lands and fell into tenancy.

Thus, as the commercial economy extended into rural villages, social divisions arose among the farmers. Tax collection became unstable, and many warriors—whose stipends, still calculated in *koku*, depended upon taxes paid by the farmers—found themselves in serious financial difficulty. Despite the general improvement of agricultural technology, productivity was uneven; and in many areas, and especially during certain eras, periodic crop failures and famines, exacerbated by excessive taxation, resulted in people starving or fleeing their villages. The abandonment of cultivated land also became conspicuous. As noted above, the samurai class had long since taken up normal residence in the cities. With the development of the urban way of life, they now incurred increasing expenses. Living on fixed incomes, many became greatly impoverished. At times, both the *bakufu* and the domains tried to suppress commercial production as a means of alleviating the suffering of their vassals; but this met with great resistance from merchants and affected the self-sufficient economy of the farmers as well. It was, in any event, a hopeless effort, given the scale of commercial development nationwide. When attempts to restrict production failed, *bakufu* and *han* administrators encouraged such production, seeking to supplement their finances by monopolizing the farmers' commercial goods and selling them themselves.

Ultimately, such rural conditions led to major outbreaks of violence. Stratification of rural villages—a growing gap between wealthy and poor farmers—tenancy, the inability of many to survive the harsh realities of commercialization, and exploitation by feudal lords forced some peasants into uprisings (*hyakushō ikki*). Even in early Edo times, there were localized demonstrations against daimyo for excessive taxation, but from the 18th century peasant protest became increasingly violent and widespread. Some uprisings were directed at local lords, some were more widespread, and some were directed not at feudal warrior overlords but at wealthy peasant landlords and village headmen who also had become exploitative. Meanwhile, economic conditions in the cities—to which frustrated peasants often fled seeking a better life—were hardly better. While many wealthy merchants enjoyed luxurious lifestyles in cooperation with warrior rulers, the city poor, driven to the edge of starvation by the rising prices of rice and other commodities, often rioted, plundering and destroying rice shops and pawnshops.

Peasant uprisings

Political reform in the *bakufu* and the *han*. The second half of the Tokugawa period is characterized by continual political reforms made by the samurai overlords in response to this ongoing economic crisis. Such reforms began with the Kyōhō Reforms instituted by the eighth shogun, Tokugawa Yoshimune (ruled 1716–45). Yoshimune swept out officials favoured by his two predecessors and appointed new officials to posts in finance and rural administration in order to increase government efficiency. In general, he reaffirmed the influence of the *fudai* daimyo, the traditional stalwart supporters of the regime, whose power had been undercut. The thrust of his reform efforts, however, came in the area of general economic policy and the *bakufu*'s own finances. As an emergency policy, Yoshimune ordered the daimyo to make rice contributions (*agemai*), which he then allotted to the *hatamoto* to supplement their stipends. More characteristic was his effort to increase tax yields by opening new lands to cultivation and revising the method of taxation. His attempt to control the falling price of rice earned him the name of "the rice shogun." But when the price of rice rose sharply in a great famine in the 1730s, the common people of Edo attacked the wholesale rice dealers who had cornered the market. This was the first such riot in Edo. Yoshimune's economic reforms enjoyed no small success. By 1744, the year before his retirement, the receipts of the *bakufu* both in total land taxes and in tax receipts reached their highest level for the entire Edo period. Yoshimune's reforms also expedited the legal process, ameliorated punishments, and were published in a collection of laws (*Kujikata osadamegaki*). For such reasons, Yoshimune was regarded as the restorer of the *bakufu*.

Under the rule of Yoshimune's son Ieshige, control of government by attendants of the shogun—which Yoshimune's strong personal rule had prevented—was revived. Chamberlains (*soba-yōnin*) who handled communications with the senior councillors (*rōjū*), gained strong powers of authority as his spokesmen when they won the shogun's confidence. One such man was Tanuma Okitsugu, who rose from chamberlain to be senior councillor under Ieshige's son, Ieharu, the 10th shogun. Tanuma and his associates accepted bribes, and he was criticized by an opposition group for corruption. But Tanuma was nonetheless an active reformer who further developed some of Yoshimune's programs. Tanuma chose not to suppress the activities of big-city merchants but rather used them to promote production; while advancing the development of the commercial economy, he sought to control it. His decision to force commercial and industrial guilds, or *kabu nakama,* into monopolistic associations and to demand licensing fees seems to have been aimed at gaining control over the circulation of commercial goods, linking the city guilds with village producers. Tanuma, too, was concerned with a monetized economy, especially the problem of money lenders. He tried—unsuccessfully—both to control the issuance of unbacked promissory notes and to issue a new silver coin, the value of which was calculated in terms of gold. He was widely criticized by the people for issuing large amounts of debased coinage that caused a rise in prices; yet it was a rational attempt to establish a gold standard in place of the confusing practice of using silver in western Japan and gold in the east.

Tanuma Okitsugu

Tanuma's rational and progressive political attitude is best revealed in his attempt to develop Ezo (present-day Hokkaido) as a bulwark against the southward advance of the Russians. Various natural disasters occurred in his time, however, and peasant protests rose to more than 50 per year during the 1780s. A great eruption of Mount Asama in 1783 was followed by a widespread famine during the Temmei era (1783–87), in which large numbers of people starved to death. An uncommon number of crop failures, fires, epidemics, and droughts reconfirmed peoples' sense of divine displeasure with the performance of the ruler. In 1787 large-scale riots threatened Edo, Ōsaka, and other major cities. Tanuma had already been dismissed as senior councillor the previous year, and Matsudaira Sadanobu, grandson of Yoshimune and the daimyo of Shirakawa domain (in modern Fukushima prefecture), was selected as his successor. But Tanuma's supporters in the *bakufu* sought to prevent Sadanobu's appointment, and for more than six months the political situation remained a complete vacuum. The outbreak of peasant violence, however, was enough to drive Tanuma's supporters from office, and Sadanobu was appointed senior councillor.

The Kansei reforms

Sadanobu is renowned as the initiator of the Kansei reforms (1789–1801). He rejected Tanuma's administration and instituted a policy of retrenchment in the spirit of Yoshimune's reforms half a century earlier. Sadanobu sought to restore morale, revive the economy, and reinvigorate the social system. He set out to reduce the high prices in the great cities and had a fund established in Edo called *shichibu tsumikin* (70 percent reserve fund); knowing that land and house rents were high in the shogun's capital because of the heavy taxes levied on its landlords, he reduced this tax and set aside 70 percent of it for relief for the poor. To relieve the hardships of the *bakufu* retainers, he took emergency measures to cancel the debts of the *hatamoto* to the Edo merchants who handled the exchange of their stipends. The farming villages, which were the foundation of the *bakuhan* system, had been devastated in the Temmei famine of the 1780s. Sadanobu encouraged officials to bring land back into cultivation and to increase the population of the villages by such measures as granting parcels of land to vagabonds. Those who had left villages for seasonal work in cities were given money to return to agricultural productivity. Infanticide and abortion were widely used as means of limiting family size, both to maximize wealth and to avoid starvation, and pregnant women were thus watched over in order to increase the farming populace—so that tax revenues from that sector would rise. But effective control of agriculture depended largely upon competent officials, and, despite heroic efforts to root out incompetence and avarice, Sadanobu had recurrent problems dealing with corrupt and ruthless local officials.

Sadanobu was a firm admirer of Chu Hsi studies, and he believed that government must be conducted on the basis of Confucian benevolent rule. In the mid-1790s, he prohibited all teachings except those of the Chu Hsi school at the Shōheikō, the *bakufu* official college headed by the Hayashi family. He even instituted a five-level examination system for promotions among *bakufu* officials who were trained at this shogunal academy.

While Sadanobu was senior councillor, a Russian envoy, Adam Laxman, landed at Nemuro in 1792 and requested trade relations. Although the *bakufu* rejected the Russian proposal, Sadanobu ordered that plans be drawn up immediately for a coastal defense system centred on Edo Bay (now called Tokyo Bay), while he himself inspected the coastline of Izu, Sagami, and Bōsō. At Sadanobu's resignation in 1793 these plans were scrapped, but the *bakufu* councillors of this era were the first to react to the threat of foreign nations advancing on Japan, which now could be heard through the wall of national seclusion. Sadanobu's reforms appear to be an overreaction to Tanuma's administration, and, whereas people at first welcomed them, antipathy gradually increased. Within the *bakufu* Sadanobu had his enemies. Some Tanuma supporters remained in *bakufu* posts through his early years; in addition, the *ōoku* (women's quarter, the shogun's harem), disliked him since he had purged some women who had become involved with Buddhist priests. Ultimately, he lost the confidence of the shogun Ienari and resigned.

In conjunction with the *bakufu* programs, reforms were carried out within the various daimyo domains. A distinctive feature of *han* reforms at this time, however, was that they tried to apply stronger regulations and control over the commercial economy of the farmers.

Ienari was restrained by Sadanobu's strict political reforms, but, when the latter left the *bakufu* council, the shogun was able to relax. Even so, Ienari was not completely free while the councillors who had supported Sadanobu's reforms were still alive. During the period 1804–31 these men died one after another, and the *bakufu* government became lax once again. Mizuno Tadaakira, a senior councillor with acute business acumen, rose to power as a personal attendant to Ienari. But he and other high officials seemed as addicted to bribery as earlier regimes, and the corruption of the *bakufu* increased considerably. On the surface things seemed peaceful, but underneath the stagnation of the feudal system became even more grave. Even in the villages of Kantō, the seat of the *bakufu,* disturbances continued apace. The *bakufu* therefore set up an office called the Kantō Torishimari-deyaku ("Supervisors of the Kantō District") to strengthen police control of the area, and it ordered the villages of Kantō to form associations to assist this office. But the impetus to reform had faded, as almost a century of *bakufu* efforts to deal effectively with vastly changed socioeconomic conditions had proved ineffective: many samurai "rulers" lived in poverty, while officially despised merchants became incredibly wealthy; the number of tenants soared as the gap between rich and poor farmers widened; commercialization proceeded far beyond the understanding as well as the control of the regime; rural and urban unrest threatened the stability of society; and now the ominous spectre of a foreign threat loomed on the horizon.

Increasing corruption

The growth of the northern problem. In the early 1800s foreign relations, which national seclusion policies had been designed to avoid, became a pressing problem for the *bakufu,* and the situation in Ezo became especially worrisome. In 1804 another Russian envoy, N.P. Rezanov, visited Japan—this time at Nagasaki, where the Dutch by law were allowed to call—to request commercial relations. The *bakufu* refused Rezanov's request, and during the next three years Russians attacked Sakhalin and the Kuril Islands. Earlier in 1804, the *bakufu* had taken eastern Ezo away from the jurisdiction of the Matsumae domain in northern Honshu and placed it under its direct control, and in 1807 the *bakufu* also took direct control of both eastern and western Ezo for defensive purposes. In 1808 the English warship *Phaeton* made an incursion on Nagasaki, and three years later the Russian naval lieutenant V.M. Golovnin landed on Kunashiri Island, where he was arrested by *bakufu* authorities. When these various incidents were resolved, peace continued for a time in the northern regions; the *bakufu* relaxed its precautions, returning all Ezo to the control of the Matsumae domain in 1821. In the south, English ships often appeared in Japanese waters after the *Phaeton* incident, and the *bakufu* failed to adopt a consistent policy. In 1825, responding to a proposal by Takahashi Kageyasu, Edo authorities promulgated the Order to Drive Away Foreign Ships (*Ikokusen uchiharairei*), which also enjoined coastal authorities to arrest or kill any foreigners who came ashore. This was also known as the *ninen nashi* or "no second thought" law. It was never fully carried out because of opposition by a number of officials, including Matsudaira Sadanobu. In 1842, upon hearing the news of China's defeat in the Opium War, the *bakufu* responded to foreign demands for the right to refuel in Japan by canceling that order and adopting the Order for the Provision of Firewood and Water (*Shinsui kyūyorei*).

Weakening of national seclusion

While attempting to preserve the iron law of seclusion to the bitter end, *bakufu* policy was thus inconsistent, driving foreign ships away at one point and treating them with leniency at others. And it proved to be utterly powerless when it was faced with the full weight of foreign pressure later in the 1840s.

New learning and thought. Underlying this weakening of the *bakuhan* political system was an ideological crisis, the result of many new movements that took place in scholarship and culture. The mid-Tokugawa period, roughly the 18th century, as discussed above, was a time of considerable unrest. Samurai leaders of *bakufu* and *han* alike sought to grapple with the disturbing fact that the great peace envisioned as resulting from policies of rigid class separation, national isolation, and agricultural self-sufficiency was being undermined by unintended economic changes released by those policies themselves. In the area of thought, the ideological foundations of Edo rule—orthodox Chu Hsi philosophy—came into question. Ironically, the ideal of "the investigation of things" inherent in Chu Hsi philosophy encouraged speculation that inevitably led to questioning Chu Hsi orthodoxy itself.

Heterodox Confucian schools. Already in the second half of the 17th century the scholars of the *kogaku* ("study of antiquity") school criticized Chu Hsi studies and advocated a return to the original ideals of Confucianism. Two of the most important thinkers articulating this view were Itō Jinsai and Ogyū Sorai. Sorai, acknowledged to be the seminal thinker of Edo times, was especially concerned with the contradictions between social theory and reality. Critical of the rise of merchants and farmers at the expense of the samurai, he tried to find a way to revive the deteriorating conditions of warriors. In his work *Seidan,* for example, Sorai insisted that the main reason for the financial distress of the warrior class in both the *bakufu* and the domains was that warriors had moved to the cities, where they were at the mercy of a monetary economy. If they would return to the villages, they could be self-sufficient once again, and other orders of society—especially the peasants—would respect them. The proper relations between the classes could thus be restored. Kogaku critics of orthodoxy were hardly alone. Various other schools of Confucianism arose, such as *setchūgaku* ("eclectic school") and *kōshōgaku* ("positivistic school"). The *bakufu* attempted to reinvigorate Chu Hsi orthodoxy by prohibiting all other schools of Confucianism in the college of the *bakufu,* but the attempt was destined to failure. Confucianism, both Chu Hsi orthodoxy and other types, now spread widely throughout the provinces, especially with the establishment of domain schools (*hankō*) for the education of the domain samurai. Thus, learning and culture arose in the domains, accompanied by a growth of scholarship with local colouring. Among such schools, the Kaitoku-dō in Ōsaka became famous as the "townspeople's university," where samurai and merchants sat together to hear lectures.

Shintō and kokugaku. The intellectual vitality of the 18th century was not limited to Confucianism. New currents also appeared in Shintō, which, often mixed with Confucianism and Buddhism, served as the ideology of popular education. The Confucian scholar Yamazaki Ansai formulated a Shintō ideology with a distinctly Confucian bent, called the Suika form of Shintō. In the later Tokugawa period popular interest in Shintō grew progressively stronger, centred especially on faith in the shrine at Ise, which became the focus of mass pilgrimages. This popularity was spurred by public lectures that explained Shintō in terms easily understood by the common people; furthermore, Watarai Nobuyoshi, Anzai, and others decoupled Shintō from its previous amalgamation with medieval Buddhism, explaining it from a Confucian perspective. Ishida Baigan developed a religious tradition called Shingaku ("Heart Learning"), which held forth a code of self-cultivation that valued performance of one's tasks; Ishida's ideas have sometimes been regarded as a Japanese version of the "Protestant ethic."

Kokugaku

Kokugaku ("national learning") also arose from a similar social background. Kamo Mabuchi focused on a study of Japan's most ancient poetry anthology, the *Man'yōshū* and other ancient writings, urging a return to ancient ways before Japan had been "defiled" by foreign ideas, such as Confucianism and Buddhism. By studying the ancient language of Japan's oldest classic, the *Kojiki* ("Records of Ancient Matters"), Mabuchi's pupil Motoori Norinaga tried to explicate Japan's ancient system of morality, called *kannagara no michi* ("way of the gods"). Another important figure in the *kokugaku* stream was Hirata Atsutane. Atsutane accepted Norinaga's explanation of Fukko ("Restoration," or "Revival") Shintō and regarded Japan as the centre of the world; as an adherent of the belief in Japan as a divine country (*shinkoku*), he strongly advocated reverence for the imperial house. Hirata's thought, along with the Confucian-inspired loyalism of the Mito school, provided the ideological underpinnings of the "Sonnō jōi" ("Revere the Emperor! Expel the Barbarians!") movement of the last years of the Tokugawa period.

Western studies. The study of modern European science, termed *yōgaku* ("Western learning") or *rangaku* ("Dutch learning"), also attracted the attention of curious scholars, especially as the regime began to lose its efficacy. A great stimulus to the concrete development of Western studies was provided by the publication, in 1774, of the *Kaitai shinsho* ("New Book of Anatomy"), a translation by Sugita Gempaku and others of an anatomical book imported from the Netherlands. Thereafter, Western studies became increasingly dynamic, focusing primarily on medicine. But as the systemic crisis grew more severe, many scholars of Western studies began to criticize the seclusion policy, arousing the ire of the *bakufu.* As consciousness of the foreign threat grew stronger, adherents of Western studies placed heavy emphasis on the study of military technology.

Other philosophers also appeared who repudiated feudal society. Andō Shōeki rejected the stratified society established by rulers as no more than a fabrication, preaching in its place a "natural society" in which all were equal. While Shōeki may be considered exceptional in the degree of his criticism of the society, others developed critical antifeudal worldviews that were directly or indirectly influenced by empirical science and Western studies. Miura Baien of Kyushu called his learning *jōrigaku* ("rational studies"); it contained a dialectical method of thought that, rejecting the fixed "way" of orthodox Neo-Confucianism, saw the world as being constantly in flux. The naturalist Hiraga Gennai agreed substantially with Tanuma Okitsugu's desire to promote the production of various goods. Hiraga gave full play to his genius by cultivating sugarcane and carrots, producing Dutch-style pottery, and surveying and developing mines in various provinces of the country.

Shiba Kōkan and Yamagata Bantō

Two other noteworthy scholars of the late 18th and early 19th century were Shiba Kōkan and Yamagata Bantō. Kōkan was widely influenced by Dutch studies and Western rationalism in general. He criticized the Tokugawa status system on the ground that the emperor and the beggar were similar human beings, thus insisting on human equality. Bantō was a noted student of the Kaitokudō. In his work *Yume no shiro* ("Instead of Dreams"), he reconstructed Japanese history in the age of gods on the basis of natural science.

Growth of popular knowledge. The common people of the Tokugawa period, both urban and rural dwellers, by the very fact of their integration into a nationwide economic system of some technological sophistication, were increasingly reared in a world of empirical knowledge. At the outset of the period only a handful of upper-class farmers or urban merchants were literate; by the end of this period, with the exception of the very lowest class, farmers were all at least partly literate. This spread of literacy was to some extent facilitated by the diffusion of temple schools (*terakoya*), the educational organs of the common people. In any case, there was a marked growth in popular knowledge over the two and half centuries of Tokugawa rule. As an example, "village conflicts" (*murakata sōdō*) became more fierce in the later part of this period, as the farmers sought to censure the improper acts of village officials and to make the village more democratic. Leadership in these conflicts was often taken by middle- and lower-class farmers.

Religious attitudes. Despite official hostility toward systems of thought and belief other than Neo-Confucianism, Buddhism retained a strong influence over the lives of the common people. For example, the medieval sects of Jōdo, Jōdo Shin, Zen, and Nichiren made striking advances during the Edo period. Besides their previous roles conducting funeral rites and other more strictly "religious" functions, they now were charged with the official state functions of registering citizens and conducting the census. As they were thus exceedingly closely connected to the daily lives of the people, the continued existence of Buddhist temples was guaranteed. More people in Edo times tended to engage in what was termed *genze riyaku—i.e.,* they prayed for happiness during their lifetime, such as for commercial prosperity or restoration of health—rather than wait for happiness after their death, as had been more common in medieval Buddhism. In response to these practical desires and needs, temples conducted various ceremonies and concocted other means to increase their income. Two of the most important such ceremonies were *kaichō* ("displaying temple treasures") and *tomitsuki. Kaichō* consisted of allowing the people to worship a Buddhist image that was normally kept concealed and not generally displayed. Gradually this ceremony came to be performed by transporting the image to other cities and villages for display. *Tomitsuki* was an officially authorized lottery, and in Edo the raffles at such temples as Yanaka Tennō, Yushima Tenjin, and the Meguro Fudō (better known as the Ryūsen Temple) were famous. Many Buddhist priests profited from these activities, and some led rather profligate private lives, providing orthodox Confucian scholars reasons for demanding that Buddhism be stamped out. Yet, despite official disapproval, it remained important in the lives of the people.

Shugendō

Various sorts of popular faiths flourished also in the cities and villages of Edo times. Shugendō, for example, was an ancient form of ascetic practice preached by itinerant monks (*yamabushi*), who offered prayers to cure illness or bring happiness. While its teachings centred on traditional Tendai and Shingon Buddhism, it also contained beliefs drawn from Shintō, religious Taoism, and elsewhere to meet the religious feelings of the people. A new faith in healing spirits arose, sparked by the view that human suffering could be cured only by those who had suffered similar hardships themselves; and in the late Tokugawa period there developed a belief in living gods (*ikigami*) who could respond to the various needs and desires of the common people and who became revered as founders (*kyōso*) of new religious sects. Among such sects were Kurozumikyō, founded by Kurozumi Munetada, Konkōkyō of Kawate Bunjirō, and Tenrikyō of Nakayama Miki, all of which remain active in present-day Japan. A peasant girl who suffered great hardship in her personal life, Nakayama became a shaman and a faith healer and attracted a widespread following. Many such people founded new religions, espousing utopian causes and leading millenarian movements; their advocacy of *yo-naoshi,* or relief of the world by social reform, had clear political overtones. In a similar manner, others were influenced by the growth of the cult of Shintō shrines, and periodic pilgrimages to Ise became popular. Not only Ise but other shrines as well became the focus of popular pilgrimages; such major deities as Inari, Hachiman, and Tenjin became associated with local gods (*ujigami*) and developed into objects of local worship. Pilgrimages could consist of groups of several hundreds of thousands of commoners. Among these masses of Shintō pilgrims were many harbouring the same social and political hopes for *yo-naoshi* expressed in the faiths of the founders of new sects.

The maturity of Edo culture. In the early 19th century the urban culture that had arisen in the 17th century reached full maturity. Supported originally by wealthy townspeople and warriors, this Edo urban culture spread widely among the urban dwellers of Japan's major cities and castle towns. In the 17th century literary and artistic production had centered in the Kyōto-Ōsaka area, but late Tokugawa culture was primarily produced in Edo. Literary styles took various forms; representative authors are Santō Kyōden in the *sharebon* (genre novel), Jippensha Ikku in the *kokkeibon* (comic novel), and Takizawa Bakin in the *yomihon* (regular novel). They examined in detail such things as the townspeople's way of life, customs, conceptions of beauty, and ways of thinking.

In the world of art, ukiyo-e reached maturity in both form and content and was unquestionably the most popular art form. Early wood-block printing had been simply in black and white, but artists had experimented with colour. *Nishiki-e,* literally "brocade pictures" (wood-block printing in many colours), was invented by Suzuki Harunobu in 1765 and entered its golden age with the prints of kabuki actors by Tshūsai Sharaku and of courtesans by Kitagawa Utamaro. In the last years of the Edo period, the masters of wood-block landscape prints, Andō Hiroshige and Katsushika Hokusai, extended the boundaries of wood-block prints far beyond the world of the pleasure quarters. While their prints show how Japanese artists had absorbed some techniques from Western art, the popularity of their works outside Japan and influence on foreign art is a measure of the sophistication Japanese culture reached in Edo times. As a result of the development of complex transportation links and market networks between city and countryside, scholarship, literature, and art not only spread to but even was produced in regional towns and villages, where crafts and products with distinctive local colouring were supported by landlords and merchants. A national culture emerged and became the foundation of a modern Japanese culture that developed after 1868.

Wood-block prints

Yet the spread of literacy and a nationwide culture could not mask contradictions in the political sphere. There were signs of stagnation and corruption in some aspects of Edo culture—a reflection of the crisis in the *bakuhan* system. The crisis had reached new levels by the 1830s. A great famine then, the result of abnormal weather conditions and resultant crop failures, lasted several years and dealt a savage blow to the impoverished villages. Both peasant uprisings and city riots over food shortages and intolerable living conditions reached unprecedented peaks. In 1836, to cite one extreme example, an uprising in Gunnai district of Kai province (Yamanashi prefecture), then under direct *bakufu* control, eventually attracted more than 50,000 participants and for a time reduced the centre of Kai to anarchy. The depth of the *bakufu*'s shock can be gauged from the fact that they sentenced 562 persons to crucifixion for their part in the uprising. Just a year later in the *bakufu*-controlled city of Ōsaka, Ōshio Heihachirō, a former city official, led a revolt aimed at overthrowing city officials and wealthy merchants and relieving the plight of the poor. Although the uprising was speedily suppressed, the *bakufu* was again shocked, incredulous that a former faithful official would lead a revolt.

At the same time that the *bakufu* was facing these serious domestic disturbances, the European powers also began to press more heavily upon Japan. The Opium War (1839–42) broke out between Ch'ing dynasty China and Britain, and foreign encroachments on Chinese territory following the British victory filled *bakufu* authorities with a sense of crisis. Tokugawa Nariaki, lord of Mito *han,* a Tokugawa collateral domain, urged the *bakufu* to institute drastic political reforms: he called the outbreak of rural and urban violence "domestic anxiety" and the pressure of the foreign powers "foreign anxiety."

THE LAST YEARS OF THE BAKUHAN

Mizuno Tadakuni

The Tempō reforms. Thus beset by crises in both domestic and foreign affairs, the chief senior councillor (*tairō*), Mizuno Tadakuni, instituted the Tempō reforms, named for the Tempō era (1830–44). Based on the earlier Kōhyō and Kansei reforms and equally conservative, Tadakuni's efforts lasted only from 1841 to 1843. He revised the regulations for the government officials and encouraged the samurai to practice frugality and diligently study the literary and martial arts. He also aimed to restore the farming villages devastated by the great famine. Stricter than earlier reformers, Tadakuni planned to force temporary residents in Edo to return to their home villages and to restrict the commercial-goods production of the farmers to make them concentrate on rice farming. He tried to lower the prices of commodities in the cities through detailed regulations on the lives of townspeople. Tadakuni further ordered the dissolution of *kabu nakama*, the merchant and artisan guilds, since he regarded them as the cause of rising commodity prices. Concerned as well with the foreign threat, he planned to reclaim the Imba Swamp (in modern Chiba prefecture) so that food supplies could easily be conveyed to nearby Edo if Edo Bay were blockaded by foreign ships. Plans for the defense of the bay also were formulated. Tadakuni also promulgated a land-requisition (*agechi*) order to bring daimyo and *hatamoto* domains surrounding Edo and Ōsaka under direct *bakufu* control: the stated object of this was the defense of Edo, but it also was designed to supplement the finances of the *bakufu*. The *agechi* order was finally withdrawn, however, in the face of fierce opposition from the daimyo, *hatamoto*, and people of the domains affected; and, as a direct result of this failure, Tadakuni was driven from power in 1845.

Tadakuni predicted that, thanks to his reforms, the Tokugawa regime would survive for another 30 years. It was, in fact, almost 30 years after his reforms (1867) that the *bakufu* was toppled by the combined forces of several *tozama* lords. During the Tempō period, administrative reforms were carried out in many of the domains, often with far more success than those of the *bakufu*. The reforms of the powerful domains in southwestern Japan—Chōshū, Satsuma, and Hizen in particular—were especially noteworthy, where middle- and lower-class samurai, motivated by a sense of loyalty to the *han* and frustrated at having been denied participation in domain administration, came forward as reformers. Replacing the previous conservative officials, these young reformers, often with a more realistic knowledge of the outside world gained through study in Edo, Nagasaki, and Kyōto, set about to strengthen their domains and expressed opinions on the national situation as well. Adopting the slogan "Fukoku kyōhei" ("Enrich the Country, Strengthen the Military"), these new officials were able to institute policies that improved domain finances and modernized their military capabilities. The way was thus gradually being prepared for the emergence of the leaders of the Meiji Restoration (1868) and of modern Japan.

Foreign rivalries

The opening of Japan. In 1845, when Abe Masahiro replaced Mizuno Tadakuni as head of the *rōjū*, there were various reactions against the Tempō reforms. Reaction against domestic reform was comparatively calm, however, and the major stumbling block facing the *bakufu* was the foreign problem. The Netherlands, the only European power trading with Japan, realized that, if Britain succeeded in forcing Japan to open the country, it would lose its monopoly; so the Dutch now planned to seize the initiative in opening Japan and thus to turn the situation to their own advantage. In 1844 the Dutch sent a diplomatic mission urging the *bakufu* to open the country, but Abe and the *bakufu* rulers refused this suggestion. Yet visits by foreign ships proliferated. In 1844, 1845, and 1846, British and French warships visited the Ryukyu Islands and Nagasaki to request commercial relations. In response, the *bakufu* in 1845 established a new office for coastal defense and various diplomatic posts. The defense system of Edo Bay also was revived, the number of domains on guard duty was increased, and new gun emplacements were built. In 1848 the *bakufu* decided not to revive the Order to Drive Away Foreign Ships, which had been rescinded during the Tempō reforms, but decided instead to continue extensive military preparations against potential attack.

Rumours had long circulated among the various Western powers that the U.S. government would send an expeditionary fleet to Japan. In 1846 Commander James Biddle of the American East Indian fleet appeared with two warships in Uraga Harbour (near Yokohama) and held consultations with *bakufu* representatives on the question of opening commercial relations. When refused by the *bakufu*, Biddle returned empty-handed. The United States, however, eagerly desired ports for fuel and provisions for its Pacific merchant and whaling ships and was not willing to give up attempts to open Japan. But the *bakufu* had for two centuries retained its political dominance through strict adherence to the policy of seclusion, and it could not muster up the resolution necessary to open the country. Opinion among the daimyo and samurai was split between seclusion and opening the country. The opening of Japan was thus postponed until the last possible moment and had to be effected unilaterally by foreign pressure, backed by massive naval strength. This pressure was initiated by the squadron of U.S. warships commanded by Commodore Matthew C. Perry that entered Uraga Bay in July 1853. (K.Ma./G.C.H.III)

Japan since 1850

THE MEIJI RESTORATION

The term restoration is commonly applied to the political changes in Japan that returned power to the imperial house in 1868. In that year the boy emperor Mutsuhito—later known by his reign name Meiji, or "Enlightened Rule"—replaced the Tokugawa *bakufu*, or shogunate, at the political centre of the nation. Although phrased in traditional terms as a restoration of imperial rule, the changes initiated during the Meiji period (1868–1912) constituted a social and political revolution that began in the late Tokugawa period and was not completed until the promulgation of the Meiji constitution in 1889.

Treaties

The fall of the Tokugawa. The arrival of Americans and Europeans in the 1850s increased domestic tensions. The *bakufu*, already weakened by an eroding economic base and ossified political structure, now found itself challenged by Western powers intent on opening Japan to trade and foreign intercourse. When the *bakufu*, despite opposition from the throne in Kyōto, signed the Treaty of Kanagawa (or Perry Convention; 1854) and the Harris treaty (1858), the shogun's claim of loyalty to the throne and his role as "subduer of barbarians" came to be questioned. To bolster his position, the shogun elicited support from the daimyo through consultation, only to discover that they were firmly xenophobic and called for the expulsion of Westerners. The growing influence of imperial loyalism, nurtured by years of peace and study, received support even within the shogunal camp from men such as Tokugawa Nariaki, the lord of Mito domain (*han*). Activists used the slogan "Sonnō jōi" ("Revere the emperor! Expel the barbarians!") not only to support the throne but also to embarrass the *bakufu*. Nariaki and his followers sought to involve the Kyōto court directly in shogunal affairs in order to establish a nationwide program of preparedness. In this Nariaki was opposed by the *bakufu*'s chief councillor (*tairō*), Ii Naosuke, who tried to steer the nation toward self-strengthening and gradual opening. But Ii's effort to restore the *bakufu* was short-lived. In the spring of 1860 he was assassinated by men from Mito and Satsuma. Ii's death inaugurated years of violence during which activist samurai used their swords against the hated "barbarians" and all who consorted with them.

By the early 1860s the Tokugawa *bakufu* found itself in a dilemma. On the one hand it had to strengthen the country against foreigners. On the other it knew that providing the economic means for self-defense meant giving up shogunal controls that kept competing lords financially weak. Activist samurai, for their part, tried to push their feudal superiors into more strongly antiforeign positions. At the same time, antiforeign acts provoked stern countermeasures and diplomatic indemnities. Most samurai soon realized that expelling foreigners by force was impossible. Foreign military superiority was demonstrated conclusively with the bombardment of Kagoshima in 1863 and Shimonoseki in 1864. Thereafter, samurai activists used their antiforeign slogans primarily to obstruct and embarrass the *bakufu.* Domestically it was forced to make antiforeign concessions to placate the loyalist camp, while foreigners were assured that it remained committed to "opening the country" and abiding by the treaties. Both sides saw it as prevaricating and ineffectual. After the arrival of the British minister Sir Harry Parkes in 1865, Great Britain, in particular, saw no reason to negotiate further with the *bakufu* and decided to deal directly with the imperial court in Kyōto.

Rise of discontent in Chōshū

Samurai in several domains also revealed their dissatisfaction with the *bakufu*'s management of national affairs. One domain in which the call for more direct action emerged was Chōshū (now part of Yamaguchi prefecture), which fired on foreign shipping in the Shimonoseki Strait in 1863. This led to bombardment of Chōshū's fortifications by Western ships in 1864 and a shogunal expedition that forced the domain to resubmit to Tokugawa authority. But many of Chōshū's samurai refused to accept this decision, and a military coup in 1864 brought to power, as the daimyo's counselors, a group of men who had originally led the radical antiforeign movement. Several of these had secretly traveled to England and were consequently no longer blindly xenophobic. Their aims were national—to overthrow the shogunate and create a new government headed by the emperor. The same men organized militia units that utilized Western training methods and arms and included nonsamurai troops. Chōshū became the centre for discontented samurai from other domains who were impatient with their leaders' caution. In 1866 Chōshū allied itself with neighbouring Satsuma, fearing a Tokugawa attempt to crush all opponents to create a centralized despotism with French help.

Again shogunal armies were sent to control Chōshū in 1866. The defeat of these troops by Chōshū forces led to further loss of power and prestige. Meanwhile, the death of the shogun Iemochi in 1866 brought to power the last shogun, Yoshinobu, who realized the pressing need for national unity. In 1867 he resigned his powers rather than risk a full-scale military confrontation with Satsuma and Chōshū, doing so in the belief that he would retain an important place in any emerging national administration. But this was not to be. Outmaneuvered by the young Meiji emperor, who succeeded to the throne in 1867, and a few court nobles who maintained close ties with Satsuma and Chōshū, the shogun faced the choice of giving up his lands, which would risk revolt from his vassals, or appearing disobedient, which would justify punitive measures against him. Yoshinobu tried to move troops against Kyōto, only to be defeated. In the wake of this defeat, Satsuma, Chōshū, and Tosa units, now the imperial army, advanced on Edo, which was surrendered without battle. While sporadic fighting continued until the summer of 1869, the Tokugawa cause was doomed. In January 1868 the principal daimyo were summoned to Kyōto to learn of the restoration of imperial rule. Later that year the emperor moved into the Tokugawa castle in Edo, and the city was renamed Tokyo ("Eastern Capital").

From feudal to modern state. The Meiji government was dominated by men from Satsuma, Chōshū, and those of the court who had sided with the emperor. They were convinced that Japan needed a unified national government to achieve military and material equality with the West. Most, like Kido Kōin and Itō Hirobumi of Chōshū and Saigō Takamori and Ōkubo Toshimichi of Satsuma, were young samurai of modest rank, but they did not represent in any sense a class interest. Indeed, their measures destroyed the samurai class. In order to gain backing for their policies, they enlisted the support of leaders from domains with which they had worked—Tosa, Saga, Echizen—and court nobles like Iwakura Tomomi and Sanjō Sanetomi. The cooperation of the impressionable young emperor was essential to these efforts.

It was believed that the West depended on constitutionalism for national unity, on industrialization for material strength, and on a well-trained military for national security. "Fukoku kyōhei" ("Enrich the country, strengthen the military") became the Meiji slogan. Knowledge was to be sought in the West, the goodwill of which was essential for revising the unequal treaties. In 1871 Iwakura Tomomi led a large number of government officials on a mission to the United States and Europe.

"Fukoku kyōhei"

Abolition of feudalism. The Meiji reformers began with measures that addressed the decentralized feudal structure to which they attributed Japan's weakness. In 1869 the lords of Satsuma, Chōshū, Tosa, and Saga were persuaded to return their lands to the throne. Others quickly followed suit. The court took steps to standardize the administration of the domains, appointing their former daimyo as governors. In 1871 the governor-daimyo were summoned to Tokyo and told that the domains were officially abolished. The 250 former domains now became 72 prefectures and three metropolitan districts, a number later reduced by one-third. In the process, most daimyo were eased out of administrative roles, and though rewarded with titles in a new European-style peerage in 1884, were effectively removed from political power.

The Meiji leaders also realized that they had to end the complex class system that had existed under feudalism. Yet, it was difficult to deal with the samurai, who numbered, with dependents, almost two million in 1868. Starting in 1869 the old hierarchy was replaced by a simpler division that established three orders: court nobles and former feudal lords became *kazoku* ("peers"); former samurai, *shizoku,* and all others (including outcast groups) now became *heimin* ("commoners"). The samurai were initially given annual pensions, but financial duress forced the conversion of these into lump-sum payments of interest-bearing but nonconvertible bonds in 1876.

Many former samurai lacked commercial experience and squandered their bonds. A national conscription system instituted in 1873 further deprived samurai of their monopoly on military service. Samurai discontent resulted in numerous revolts. An uprising in Chōshū expressed dissatisfaction with administrative measures that deprived the samurai of their status and income. In Saga, samurai called for a foreign war to provide employment for their class. The last, and by far the greatest, revolt came in Satsuma in 1877. This rebellion was led by the restoration hero Saigō Takamori and lasted six months. The imperial government's conscript levies were hard-pressed to defeat Saigō, but in the end they prevailed. In this, as in the other revolts, issues were localized, and the loyalties of most Satsuma men in the central government remained with the imperial cause.

Land surveys

Land surveys were begun in 1873 to determine the amount and value of land based on average rice yields in recent years, and a monetary tax of 3 percent of land value was established. The same surveys led to certificates of land ownership for farmers, who were released from feudal controls. The land measures involved basic changes, and there was widespread confusion and uncertainty among farmers that expressed itself in the form of short-lived revolts and demonstrations. But the establishment of private ownership, and measures to promote new technology, fertilizers, and seeds, produced a rise in agricultural output.

Although it was hard-pressed for money, the government initiated a program of industrialization, which was seen as essential for national strength. Except for military industries and strategic communications, this program was largely in private hands, although the government set up pilot plants to provide encouragement. Trade and manufacturing benefited from a growing national market and legal security, but the unequal treaties enacted with foreign powers made it impossible to protect industries with tariffs until 1911.

In the 1880s fear of excessive inflation led the government to sell its remaining plants to private investors—usually individuals with close ties to those in power. As a result, a small group of men came to dominate many industries. Collectively they became known as the *zaibatsu,* or financial cliques. With great opportunities and few competitors, *zaibatsu* firms came to dominate enterprise after enterprise. Sharing a similar vision for the country, these men maintained close ties to the government leadership.

Equally important for building a modern state was the development of national identity. True national unity required the propagation of new loyalties among the general populace and the transformation of powerless and inarticulate peasants into citizens of a centralized state. The use of religion and ideology was vital to this process. Early Meiji policy, therefore, elevated Shintō to the highest position in the new religious hierarchy, replacing Buddhism with a cult of national deities that supported the throne. Christianity was reluctantly legalized in 1873, but, while important for some intellectuals, it was treated with suspicion by many in the government. The challenge remained how to use traditional values without risking foreign condemnation that the government was forcing a state religion upon the Japanese. By the 1890s the education system provided the ideal vehicle to inculcate the new ideological orientation. A system of universal education had been announced in 1872. For a time its organization and philosophy were Western, but during the 1880s a new emphasis on ethics emerged as the government tried to counter excessive Westernization and followed European ideas on nationalist education. In 1890 the Imperial Rescript on Education (Kyōiku Chokugo) laid out the lines of Confucian and Shintō ideology, which constituted the moral content of later Japanese education. Thus, loyalty to the emperor, who was hedged about with Confucian teachings and Shintō reverence, became the centre of a citizen's ideology. To avoid charges of indoctrination, the state distinguished between this secular cult and actual religion, permitting "religious freedom" while requiring a form of worship as the patriotic duty of all Japanese.

Constitutional movement. Many Japanese believed that constitutions provided the unity that gave Western nations their strength. The Meiji leaders therefore sought to transform Japan in this direction. In 1868 the government experimented with a two-chamber house, which proved unworkable. Meanwhile, the emperor's charter oath of April 1868 committed the government to establishing "deliberative assemblies" and "public discussion," to a worldwide search for knowledge, to the abrogation of past customs, and to the pursuit by all Japanese of their individual callings.

Echoing the government's call for greater participation were voices from below. Village leaders, who had benefited from the commercialization of agriculture in the late Tokugawa period, wanted a more participatory system that could reflect their emerging bourgeois interests. Former samurai realized that a parliamentary system might allow them to recoup their lost positions. Samurai interest was sparked by a split in the government's inner circle over a proposed Korean invasion in 1873. At odds with Iwakura and Ōkubo, who insisted on domestic reform over risky foreign ventures, Itagaki Taisuke and several fellow samurai from Tosa and Saga left the government in protest, calling for a popularly elected assembly so that future decisions might reflect the will of the people—by which they largely meant the former samurai. Starting with self-help samurai organizations, Itagaki expanded his movement for "freedom and popular rights" to include other groups. In 1881 he organized the Liberal Party (Jiyūtō), whose members were largely wealthy farmers.

Itagaki Taisuke

Under these circumstances, the emperor requested the advice of his ministers on constitutional matters. Ōkuma Shigenobu, a leader from Saga, submitted a relatively liberal constitutional draft in 1881, which he published without official approval. He also revealed sensational evidence of corruption in the disposal of government assets in Hokkaido. For this he was forced out of the government's inner circle. Ōkuma organized the Progressive Party (Kaishintō) in 1882 to further his British-based constitutional ideals, which attracted considerable support among urban business and journalistic communities.

The clamour of 1881 resulted in an imperial promise of a constitution by 1889. The constitution was drafted behind the scenes by a commission headed by Itō Hirobumi and aided by the German constitutional scholar Hermann Roesler. The period of its drafting coincided with an era of great economic distress in the countryside. This provided an environment in which party agitation could easily kindle direct action and violence, and several incidents of this type led to severe government reprisals and increased police controls and press restrictions. Village leaders, confronted by unruly members of their community whose land faced imminent foreclosure, became less inclined to support liberal ideas. Consequently, the parties decided to dissolve temporarily in 1884. In the interim Itagaki traveled to Europe and returned convinced more than ever of the need for national unity in the face of Western condescension.

Itō also traveled to Europe as part of the work to prepare the new constitution. In Germany he found an appropriate balance of imperial power and constitutional forms that seemed to offer modernity without sacrificing effective control. To balance a popularly elected lower house, Itō established a new European-style peerage in 1884. Government leaders, military commanders, and former daimyo were given titles and readied for future seats in a house of peers. A cabinet system, in which ministers were directly appointed by the emperor, was installed in 1885, and a Privy Council, designed to judge and safeguard the constitution, was set up in 1888.

The Meiji constitution

The constitution was formally promulgated in 1889, and elections for the lower house were held to prepare for the initial Diet (Kokkai), which met in 1890. The constitution took the form of a gracious gift from the sovereign to his people, and it could be amended only upon imperial initiative. Its provisions were couched in general terms. Rights and liberties were granted "except as regulated by law." If the Diet refused to approve a budget, the one from the previous year could be followed. The emperor was "sacred and inviolable"; he commanded the armies, made war and peace, and dissolved the lower house at will. Effective power thus lay with the executive, which could claim to represent the imperial will.

Despite its antidemocratic features, the constitution provided a much greater arena for dissent and debate than had previously existed. The lower house could initiate legislation. Private property was inviolate, and freedoms, though subject to legislation, were greater than before. Even military budgets required Diet approval for increases. Initially, a tax qualification of 15 yen limited the electorate to about 500,000; this was lowered in 1900 and 1920, and in 1925 universal manhood suffrage came into effect. The government leaders found it harder to control the lower house than initially anticipated, and party leaders found it advantageous, at times, to cooperate with the oligarchs. The constitution thus basically redefined politics for both sides. It also ended the revolutionary phase of the Meiji Restoration. With the new institutions in place, the oligarchs withdrew from power and were content to maintain and conserve the ideological and political institutions they had created through their roles as elder statesmen (*genrō*).

THE EMERGENCE OF IMPERIAL JAPAN

Foreign affairs. Achieving equality with the West was one of the primary goals of the Meiji leaders. Treaty reform, designed to end the foreigners' judicial and economic privileges provided by extraterritoriality and fixed customs duties was sought as early as 1871 when the Iwakura mission went to the United States and Europe. The Western powers insisted, however, that they could not revise the treaties until Japanese legal institutions were reformed along European and American lines. Efforts to reach a compromise settlement in the 1880s were rejected by the press and opposition groups in Japan. It was not until 1894, therefore, that treaty provisions for extraterritoriality were formally changed.

During the first half of the Meiji period, Asian relations were seen as less important than domestic development. In 1874 a punitive expedition was launched against Formosa (Taiwan) to chastise the aborigines for murdering Ryukyuan fishermen. This lent support to Japanese claims to the Ryukyu Islands, which had been under Satsuma influence in Tokugawa times. Despite Chinese protests, the Ryukyus were incorporated into Japan in 1879. Meanwhile, calls for an aggressive foreign policy in Korea, aired by Japanese nationalists and some liberals, were steadily rejected by the Meiji leaders. At the same time, China became increasingly concerned about expanding Japanese influence in Korea, which China still viewed as a tributary state. Incidents on the peninsula in 1882 and 1884 that might have involved China and Japan in war were settled by compromise, and in 1885 China and Japan agreed that neither would send troops to Korea without first informing the other.

The Sino-Japanese War. By the early 1890s Chinese influence in Korea had increased. In 1894 Korea requested Chinese assistance in putting down a local rebellion. When the Chinese notified Tokyo of this, Japan quickly rushed troops to Korea. With the rebellion crushed, neither side withdrew. The Sino-Japanese War formally erupted in July 1894. Japanese forces proved to be superior on both land and sea, and, with the loss of its northern fleet, China sued for peace.

Treaty of Shimonoseki

The peace treaty negotiated at Shimonoseki was formally signed on April 17, 1895; both sides recognized the independence of Korea, and China ceded to Japan Formosa, the Pescadores Islands, and the Liaotung Peninsula, granted Japan all rights enjoyed by European powers, and made significant economic concessions. A commercial treaty giving Japan special tax exemptions and other trade and manufacturing privileges was signed in 1896. Japan thus marked its own emancipation from the unequal treaties by imposing even harsher terms on its neighbour. Meanwhile, France, Russia, and Germany were not willing to endorse Japanese gains and forced the return of the Liaotung Peninsula to China. Insult was added to injury when Russia leased the same territory with its important naval base, Port Arthur (now Lü-shun), from China in 1898. The war thus demonstrated that the Japanese could not maintain Asian military victories without Western sufferance. At the same time, it proved a tremendous source of prestige for Japan and brought the government much internal support; it also strengthened the hand of the military in national affairs.

The Russo-Japanese War. Reluctant to accept Japanese leadership, Korea instead sought Russia's help. During the Boxer Rebellion (1900) in China, Japanese troops played a major part in the allied expedition to rescue foreign nationals in Peking, but Russia occupied southern Manchuria, thereby strengthening its links with Korea. Realizing the need for protection against multiple European enemies, the Japanese began talks with England that led to the Anglo-Japanese Alliance (1902). In this pact both nations agreed to aid the other in the event of an attack by two or more powers but remain neutral if the other went to war with a single enemy. Backed by Britain, Tokyo was prepared to take a firmer stand against Russian advances in Manchuria and Korea. In 1904 Japanese ships attacked the Russian fleet at Port Arthur without warning. In the Russo-Japanese War (1904–05) that followed, Japanese arms were everywhere successful; the most spectacular victory occurred in the Tsushima Strait, where the ships of Admiral Tōgō Heihachirō destroyed the Russian Baltic fleet. But the war was extremely costly in Japanese lives and treasure, and Japan was relieved when U.S. President Theodore Roosevelt offered to mediate a peace settlement. The Treaty of Portsmouth, signed on Sept. 5, 1905, gave Japan primacy in Korea, and Russia granted to Japan its economic and political interests in southern Manchuria, including the Liaotung Peninsula. Russia also ceded to Japan the southern half of the island of Sakhalin. The victory over Russia altered the balance of power in East Asia, and it encouraged nationalist movements in India and the Middle East.

Japanese expansionism. After the conclusion of the war, Japanese leaders gained a free hand in Korea. Itō Hirobumi, sent to Korea as resident general, forced through treaties that gave Korea little more than protectorate status and ordered the abdication of the Korean king. Itō's assassination in 1909 led to Korea's annexation by Japan the following year. Korean liberties and resistance were crushed. By 1912, when the Meiji emperor died, Japan had not only achieved equality with the West but also had become the strongest imperialist power in East Asia.

During World War I Japan fought on the Allied side but limited its activities to seizing German possessions in China and the Pacific.

The Twenty-one Demands

When China sought the return of former German holdings in Shantung province, Japan responded with the so-called Twenty-one Demands, issued in 1915, that tried to pressure China into widespread concessions ranging from extended leases in Manchuria and joint control of China's coal and iron resources to policy matters regarding harbours and the policing of Chinese cities. While giving in on a number of specific issues, the Chinese resisted the most extreme Japanese demands that would have turned China into a Japanese ward. Despite its economic gains, Japan's World War I China policy left behind a legacy of ill feeling and distrust, both in China and in the West. The rapaciousness of Japanese demands and China's chagrin at its failure to recover its losses in the Treaty of Versailles (1919) cost Japan any hope of Chinese friendship. Subsequent Japanese sponsorship of corrupt warlord regimes in Manchuria and North China helped to confirm the anti-Japanese nature of modern Chinese nationalism.

The part played by Japan in the Allied intervention in Siberia following the Bolshevik Revolution of 1918 caused further concerns about Japanese expansion. One of the principal reasons for the disarmament conference held in Washington, D.C., in 1921–22, was to reduce Japanese influence. A network of treaties was designed to place restraints on Japanese ambitions while guaranteeing Japanese security. These treaties included a Four-Power Pact, between Japan, Great Britain, the United States, and France, that replaced the Anglo-Japanese Alliance, and a Five-Power Naval Limitation Treaty (with Italy) that set limits for battleships at a ratio of five for Great Britain and the United States to three for Japan. An agreement on the fortification of Pacific island bases was intended to assure Japan of security in its home waters. Finally a Nine-Power Pact would, it was hoped, protect China from further unilateral demands. Japan subsequently agreed to retire from Shantung, and, shortly thereafter, Japanese armies withdrew from Siberia and northern Sakhalin. In 1925 a treaty with the Soviet Union extended recognition to the U.S.S.R. and ended active hostilities.

Constitutional government. The inauguration of parliament in 1890 was accompanied by a vigorous and often obstreperous opposition in the lower house, and it was only a general determination to convince Western skeptics that constitutional government could work in Japan that forced party and government leaders to cooperate. The first cabinets, led by Yamagata Aritomo, Matsukata Masayoshi, and Itō, maintained the principle that the government, which represented the emperor, must be aloof from parties and that the lower house should approve government requests. This policy failed because the parties tried to increase their power and patronage and therefore sought cabinets responsible to the Diet. Only the Sino-Japanese War produced the kind of unity the constitution's makers had envisaged. Thereafter, the oligarchs formed alliances with the two parties, usually exchanging cabinet seats for support in the lower house. These arrangements proved unsatisfactory, however, when party leaders raised their sights.

Constitutional Party

In 1898 Itagaki and Ōkuma combined forces to form a single party, the Constitutional Party (Kenseitō), and were allowed to form a government. But their alliance was brittle as long-standing animosities and jealousies enabled antiparty forces among the bureaucracy and oligarchy to force their resignation within a few months.

A discernible division developed among the dwindling group of Meiji leaders. Yamagata Aritomo dominated the army and much of the bureaucracy. In power for two years after the Kenseitō cabinet, he strengthened legal and institutional safeguards against rule by political parties and secured an imperial ordinance that service ministers should be career officers on active duty; this gave the army and navy power to break cabinets. Meanwhile, Itō Hirobumi endorsed the party trend by forming the Friends of Constitutional Government Party (Rikken Seiyūkai) in 1900, which enlisted most of the former followers of Itagaki's Jiyūtō.

After 1901 both Itō and Yamagata retired from active participation in politics, and until 1913 cabinets were led by their protégés Saionji Kimmochi and Katsura Tarō. Basic decisions on politics and policy, however, continued to be made by the elder statesmen, who advised the emperor on all important matters and selected prime ministers by rotating power between the two principal factions.

With the death or enfeeblement of the first generation of oligarchs, the pattern of political manipulation changed. No subsequent group could match the prestige of the Meiji leaders. The Meiji emperor died in 1912 and was succeeded by a son who took the reign name Taishō ("Great Righteousness"; reigned 1912–26); but mental illness prevented him from approximating his father's fame. The growing prestige and power of businessmen found expression in their control of the political parties and resulted in an increasing role for professional party politicians. The *genrō*'s last attempt to seat Katsura in 1912 ended in failure, while his successor, Admiral Yamamoto Gonnohyōe, was discredited by scandals in naval procurement. Ōkuma Shigenobu emerged from retirement to head a cabinet during World War I and was succeeded by a military cabinet under General Terauchi Masatake. In 1918, however, discontent with Terauchi's reactionary posture and administrative incompetence combined with the rising power of the party professionals to bring about the appointment of Hara Takashi (Hara Kei) as prime minister. Hara was the first nontitled person to hold that office, and his appointment marked the first party cabinet. His assassination in 1921 cut short his cautious efforts to rein in military and bureaucratic power and extend the franchise. After several short-lived cabinets, a successful party cabinet was organized by Katō Takaaki in 1924. The army was reduced in size, moderate social legislation was enacted, and universal manhood suffrage extended the franchise to some 14 million voters. Meanwhile, Japan avoided stronger involvement in the civil war in China and pursued a conciliatory course with the Soviet Union, despite demands from nationalists, who utilized alleged outrages in China and the discriminatory U.S. Immigration Act of 1924 to warn of the futility of cooperating with Western nations.

Influence of the bureaucrats

But, as the parties grew in power, they tended to look to bureaucrats for leadership. The businessmen who supported the parties and the bureaucrats who led them shared a fear of the social movements that followed industrialization and the importation of foreign ideas. A growing labour movement already had been checked by a special police law introduced in 1900. This was strengthened under Katō in 1925. Conservative anxieties mounted after the Japan Communist Party (JCP) was organized in 1922. Under the Meiji constitution, party cabinets had to make peace with the military, the House of Peers, and the conservatives close to the throne. Therefore, they needed to work out their ideas for reform with utmost caution. The Diet often found itself virtually powerless, which led to disorder and corruption that did little to win popular support for representative government.

Social change. Social and intellectual changes taking place in Japan were closely related to the growth and development of industry. After the Treaty of Shimonoseki the government used the Chinese indemnity to subsidize the Yawata Iron and Steel Works in northern Kyushu, which greatly expanded Japan's heavy industrial sector. At the same time, textile and other consumer-goods industries expanded to meet Japanese needs and to earn credits required for the import of raw materials. Heavy industry was encouraged by government-controlled banks, which provided needed capital. Strategic industries, notably steel and the principal rail lines, were in state hands, but most new growth was in the private sector.

By 1900 Japan's population had expanded to nearly 45 million from a late Tokugawa base of about 30 million. Increasing numbers of Japanese were attracted to urban industrial centres. At the same time, domestic food production was hard-pressed to stay abreast of population increases. Agricultural productivity, after early improvements, slowed and stagnated, and it became necessary to import food.

The enlarged urban population produced movements of social inquiry and protest. In 1895 the industrial labour force numbered about 400,000, the majority of which were women employed in the textile mills. Several efforts to organize socialist movements met with police repression. Peace-preservation laws were passed in 1900 and 1925 to inhibit labour organization, and in 1928 it became a capital crime to agitate against private property or the Japanese "national polity" (*kokutai*). The socialist movement gained strength after World War I, but its program was often theoretical and doctrinaire, and its leaders found it difficult to make contact with workers. Meanwhile, the increasing confidence and power of management came to influence, and at times control, the political parties.

In the countryside the principal reflection of Japan's growing involvement in the world economy was the increased production of silkworms, which augmented farm income. At the same time, rural Japan provided the bulk of the labourers for the new industries, and daughters from farming families were found in many textile plants.

Rural problems

But the early 20th century was not a time of agricultural prosperity. Colonial competition tended to depress domestic agricultural prices. Farmers also were handicapped by growing fragmentation of holdings and increasing tenancy. Government efforts to address the situation resulted in little more than a law in 1924 that called for mediation of landlord-tenant disputes. A financial panic in 1927 aggravated rural conditions and indebtedness, even before the collapse of the American silk market in 1929 spelled disaster for farmers and workers alike.

The most lasting social changes were found in the great metropolitan centres, where a growing labour force and new middle class were concentrated. The Tokyo-Yokohama area was devastated by the great Kantō earthquake of September 1923, and the region's reconstruction as a modern metropolis symbolized the growth of the urban society. Cultural interests during and after World War I were uniformly international and largely American in inspiration. Western music, dancing, and sports became popular, and rising urban living standards and expectations produced the need for more and better higher education. The participation of women in office work and other new occupations, and the rise of a feminist movement, however unsuccessful, marked the beginning of changes in the family system.

The educated class grew in size and vigour. Currents of thought included Western-style democracy and the new radicalism of the Soviet Union; the Marxist influence went far beyond the ranks of the struggling Communist Party—which, in any event, was soon crushed by the police. Political liberalism was championed by the educator and politician Yoshino Sakuzō, and Minobe Tatsukichi, a distinguished constitutional theorist, introduced the idea that the emperor was an organ of the state and not the sole source of sovereignty. Such men faced sharp criticism and, in time, were forced to resign their positions, but they had great influence.

The base for these new currents, however, was precarious. Politically and institutionally, no advances—beyond the suffrage act of 1925—were made, while the peace-preservation laws of 1928 established a special police corps to ferret out "dangerous thoughts." The economic well-being of the urban classes depended on the continued expansion of international trade. When the worldwide financial collapse at the end of the decade wrecked Japan's foreign markets and removed the possibility of villagers augmenting their meagre incomes from rice farming with silk production, and when the venality, irresponsibility, and occasional corruption of Diet representatives was contrasted to the poverty found in many parts of Japan, numerous Japanese were prepared to listen to charges that the political-party government, dominated by selfish *zaibatsu* interests, had neglected Japan's markets in China, imperiled morality and decency at home, and allowed subversive trends to flourish, while politicians reaped personal fortunes.

THE RISE OF THE MILITARISTS

The notion that expansion through military conquest would solve Japan's economic problems gained currency during the Great Depression of the 1930s. It was argued that the rapid growth of Japan's population—which stood

Japanese expansion in the late 19th and 20th centuries.

at close to 65 million in 1930—necessitated large food imports. To sustain such imports, Japan had to be able to export. Western tariffs limited exports, while discriminatory legislation in many countries and anti-Japanese racism served as barriers to emigration. Chinese and Japanese efforts to secure racial equality in the League of Nations covenant had been rejected by Western statesmen. Thus, it was argued, Japan had no recourse but to use force.

Military's distrust of party government

The weakening of party government. To these economic and racial arguments was added the military's distrust of party government. The Washington Conference had allowed a smaller ratio of naval strength than the navy desired, while the government of Prime Minister Hamaguchi Osachi in 1930 had accepted the London Naval Conference's limits on heavy cruisers over military objections. In 1925 Katō Takaaki had cut the army by four divisions. Many military men objected to the restraint shown by Japan toward the Chinese Nationalists' northern expedition of 1926 and 1927 and wanted Japan to take a harder line in China. Under Prime Minister Tanaka Giichi the Seiyūkai cabinet reversed earlier policy by intervening in Shantung in 1927 and 1928. But Tanaka was replaced by Hamaguchi in 1929, and under his cabinet the policy of moderation was restored.

While many military leaders chafed under the restrictions that civilian governments placed upon them, they still retained considerable power. It would be wrong to attribute such resentment to all, or even most, of the high command, but enough army officers held such views to become a locus for dissatisfaction among other groups in Japanese society.

A number of rightist organizations existed that were dedicated to the theme of internal purity and external expansion. These sought to preserve what they thought was unique in the Japanese spirit and fought against excessive Western influence. Some originated in the Meiji period, when nationalists had felt obliged to work for a "fundamental settlement" of differences with Russia. Most, like the Black Dragon Society (Kokuryūkai), combined continental adventurism and a strong nationalist stance with opposition to party government, big business, acculturation, and Westernization. By allying with other rightists, they alternately terrorized and intimidated their presumed opponents. A number of business leaders and political figures were killed, and the assassins' success in publicizing and dramatizing the virtues they claimed to embody had a considerable impact on the troubled 1930s. It is clear, however, that the terrorists never had as much influence as they claimed or as the West believed.

The principal force against parliamentary government was provided by junior military officers, who were largely from rural backgrounds. Distrustful of their senior leaders, ignorant of political economy, and contemptuous of the urban luxuries of politicians, such officers were ready marks for rightist theorists. Many of them had goals that were national-socialist in character. Kita Ikki, a former socialist and one-time member of the Black Dragon Society, contended that the Meiji constitution should be suspended in favour of a revolutionary regime advised by "national patriots" and headed by a military government, which should nationalize large properties, limit wealth, end party government and the peerage, and prepare to take the leadership of a revolutionary Asia. Kita helped persuade a number of young officers to take part in the violence of the 1930s with the hope of achieving these ends.

Aggression in Manchuria. The Kwantung Army, which occupied the Kwantung (Liaotung) Peninsula and patrolled the South Manchurian Railway zone, included officers who were keenly aware of Japan's continental interests and were prepared to take steps to further them. They hoped to place the civilian government in an untenable position and to force its hand. The Tokyo terrorists similarly sought to change foreign as well as domestic policies. The pattern of direct action in Manchuria began with the murder in 1928 of Chang Tso-lin, the warlord ruler of Manchuria. The action, though not authorized by the Tanaka government, helped bring about its fall. Neither the cabinet nor the Diet dared to investigate and punish those responsible. This convinced extremist officers that their lofty motives would make retribution impossible. The succeeding government of Prime Minister Hamaguchi sought to curtail military activists and their powers. The next plots were aimed at replacing civilian rule, and Hamaguchi was mortally wounded by an assassin in 1930.

The Mukden Incident

On Sept. 18, 1931, came the Mukden (or Manchurian) Incident, which launched Japanese aggression in East Asia. A Kwantung Army charge that Chinese soldiers had tried to bomb a South Manchurian Railway train (which arrived at its destination safely) resulted in a speedy and unauthorized capture of Mukden (now Shen-yang), followed by the occupation of all Manchuria. The civilian government in Tokyo could not stop the army, and even army headquarters was not always in full control of the field commanders. Prime Minister Wakatsuki Reijirō gave way in December 1931 to Inukai Tsuyoshi. On May 15, 1932, naval officers took the lead in a terrorist attack in Tokyo that cost Inukai his life but failed to secure a proclamation of martial law. The army now announced that it would accept no party cabinet. To forestall its desire for power, the last *genrō,* Saionji, suggested retired Admiral Saitō Makoto as prime minister. Plotting continued, culminating in a revolt of a regiment about to leave for Manchuria. On Feb. 26, 1936, several outstanding statesmen (including Saitō) were murdered; Prime Minister Okada Keisuke escaped when the assassins mistakenly shot his brother-in-law. For more than three days the rebel units held much of downtown Tokyo. When the revolt was put down on February 29, the ringleaders were quickly arrested and executed. Within the army, the influence of the young extremists now gave way to more conservative officers and generals who were less concerned with domestic reform, while sharing many of the foreign-policy goals of the young fanatics.

The only possible source of prestige sufficient to thwart the military lay with the throne. But the senior statesmen were cautious lest they imperil the imperial institution itself. The young emperor Hirohito had been enthroned in 1926, taking as his reign name Shōwa ("Enlightened Peace"). His outlook was more progressive than that of his predecessors; he had traveled in the West, and his interests lay in marine biology. Those close to the throne feared that a strong stand by the emperor would only widen the search for victims and could lead to his dethronement. As international criticism of Japan's aggression grew, many Japanese rallied to support the army.

The road to World War II. Each advance by the military extremists gained them new concessions from the moderate elements in the government and brought greater foreign hostility and distrust. Rather than oppose the military, the government agreed to reconstitute Manchuria as an "independent" state, Manchukuo. The last Manchu emperor of China, P'u-i, was declared regent and later enthroned as emperor in 1934. Actual control lay with the Kwantung Army, however; all key positions were held by Japanese, with surface authority vested in cooperative Chinese and Manchu. A League of Nations committee recommended in October 1932 that Japanese troops be withdrawn, Chinese sovereignty restored, and a large measure of autonomy granted to Manchuria. The League called upon member states to withhold recognition from the new puppet state. Japan's response was to formally withdraw from the world body in 1933. Thereafter, Japan poured technicians and capital into Manchukuo, exploiting its rich resources to establish the base for the heavy-industry complex that was to undergird its "new order" in East Asia.

Events in China. In northern China, boundary areas were consolidated in order to enlarge Japan's economic sphere. In early 1932 the Japanese navy precipitated an incident at Shanghai in order to end a boycott of Japanese goods; but Japan was not yet prepared to challenge other powers for control of central China, and a League of Nations commission arranged terms for a withdrawal. By 1934, however, Japan had made it clear that it would brook no interference in its China policy and that Chinese attempts to procure technical or military assistance elsewhere would bring Japanese opposition.

Military revolt in Tokyo

Further external ambitions had to wait, however, for the resolution of domestic crises. The military revolt in Tokyo in February 1936 marked the high point of extremist action. In its wake power shifted to the military conservatives. Moreover, the finance minister Takahashi Korekiyo, whose policies had brought Japan out of its economic depression, was killed, and his opposition to further inflationary spending was thus stilled. In politics, the confrontation between the parties and the army continued. Efforts to find a leader who could represent both military and civilian interests led to the appointment to the premiership of the popular but ineffective Konoe Fumimaro, scion of an ancient court family, in 1937. Meanwhile, in China the Nationalist leader Chiang Kai-shek had been kidnapped in the Sian Incident in December 1936, and this led to an anti-Japanese united front by Nationalists and Communists. Domestic politics revealed, moreover, that the Japanese people were not yet prepared to renounce their parliamentary system. In the spring of 1937, general elections showed startling gains for the new Social Mass (or Social Masses) Party (Shakai Taishūtō), which received 36 out of 466 seats, and a heavy majority of the remainder went to the Seiyūkai and Minseitō, which had combined forces against the government and its policies. The time seemed ready for new efforts by civilian leaders, but in the field the armies preempted them.

On July 7, 1937, Japanese troops engaged Chinese units at the Marco Polo Bridge near Peking, leading to warfare between China and Japan. Japanese armies took Nanking, Han-k'ou (Hankow), and Canton despite vigorous Chinese resistance; Nanking was brutally pillaged by Japanese troops. To the north, Inner Mongolia and China's northern provinces were invaded. On discovering that the Nationalist government, which had retreated up the Yangtze to Chungking, refused to compromise, the Japanese installed a more cooperative regime at Nanking in 1940.

Foreign relations. In November 1936 Japan signed the Anti-Comintern Pact with Germany and later with Italy. This was replaced by the Tripartite Pact in September 1940, which recognized Japan as the leader of a new order in Asia; Japan, Germany, and Italy agreed to assist each other if they were attacked by any additional power not yet at war with them. The intended target was the United States, since the Soviets and Nazis had already signed a nonaggression pact in 1939, and the Soviets were invited to join the new agreement later in 1940.

Japanese relations with the Soviet Union were considerably less cordial than those with Germany. The Soviets consented, however, to sell the Chinese Eastern Railway to the South Manchurian Railway in 1935, thereby strengthening Manchukuo. In 1937 the Soviet Union signed a nonaggression pact with China, and in 1938 and 1939 Soviet and Japanese armies tested each other in two full-scale battles along the border of Manchukuo. But in April 1941 a neutrality pact was signed with the Soviet Union, with Germany acting as intermediary.

Ties with Germany

Japanese-German ties were never close or effective. Both parties were limited in their cooperation by distance, distrust, and claims of racial superiority. The Japanese were uninformed about Nazi plans for attacking the Soviet Union, and the Germans were not told of Japan's plans to attack Pearl Harbor in Hawaii. Nor, despite formal statements of rapport, did Japan's state structure approach the totalitarianism of the Nazis. A national-mobilization law (1938) gave the Konoe government sweeping economic and political powers, and in 1940, under the second Konoe cabinet, the Imperial Rule Assistance Association was established to merge the political parties into one central organization; yet, the institutional structure of the Meiji constitution was never altered, and the wartime governments never achieved full control over interservice competition. The Imperial Rule Assistance Association failed to mobilize all segments of national life around a leader. The emperor remained a symbol, albeit an increasingly military one, and no führer could compete without endangering the national polity.

Japan's relations with the democratic powers deteriorated steadily. The United States and Great Britain did what they could to assist the Chinese Nationalist cause. The Burma Road into southern China permitted the transport of minimal supplies to Nationalist forces. Constant Japanese efforts to close this route led to further tensions between Great Britain and Japan. Anti-Japanese feeling strengthened in the United States, especially after the sinking of a U.S. gunboat in the Yangtze River in 1937. In 1939 U.S. Secretary of State Cordell Hull renounced the 1911 treaty of commerce with Japan, and thus embargoes became possible in 1940. President Franklin Roosevelt's efforts to rally public opinion against aggressors included efforts to stop Japan, but, even after war broke out in Europe in 1939, American public opinion rejected involvement abroad.

Prologue to war. The European war presented the Japanese with tempting opportunities. After the Nazi attack on Russia in 1941, the Japanese were torn between German urgings to join the war against the Soviets and their natural inclination to seek richer prizes from the European colonial territories to the south. In 1940 Japan occupied northern Indochina in an attempt to block access to supplies for the Chinese Nationalists, and in July 1941 it announced a joint protectorate with Vichy France over the whole colony. This opened the way for further moves into Southeast Asia.

The United States reacted to the occupation of Indochina by freezing Japanese assets and embargoing oil. The Japanese now faced the choices of either withdrawing from Indochina, and possibly China, or seizing the sources of oil production in the Dutch East Indies. Negotiations with Washington were initiated by the second Konoe cabinet. Konoe was willing to withdraw from Indochina, and he sought a personal meeting with Roosevelt, hoping that any U.S. concessions or favours would strengthen his hand against the military. But the State Department refused to agree to such a meeting without prior Japanese concessions. Having failed in his negotiations, Konoe resigned in October 1941 and was immediately succeeded by his war minister, General Tōjō Hideki. Meanwhile, Secretary of State Hull rejected Japan's "final offer": Japan would withdraw from Indochina after China had come to terms in return for U.S. promises to resume oil shipments, cease aid to China, and unfreeze Japanese assets. With Japan's decision for war made, the negotiators received instructions to continue to negotiate, but preparations for the opening strike against the U.S. fleet at Pearl Harbor were already in motion. Japan's war aims were to establish a "new order in East Asia," built on a "coprosperity" concept that placed Japan at the centre of an economic bloc consisting of Manchuria, Korea, and North China that would draw on the raw materials of the rich colonies of Southeast Asia, while inspiring these to friendship and alliance by destroying their previous masters. In practice, "East Asia for the Asiatics," the slogan that headed the campaign, came to mean "East Asia for Japan."

Attack on Pearl Harbor

World War II and defeat. The attack on Pearl Harbor (Dec. 7 [Dec. 8 in Japan], 1941) achieved complete surprise and success. It also unified American opinion and determination to see the war through to a successful conclusion. The Japanese had expected that, once they fortified their new holdings, a reconquest would be so expensive in lives and treasure that it would discourage the "soft" democracies. Instead, the U.S. fleet was rebuilt with astonishing speed, and the chain of defenses was breached before the riches of the newly conquered territories could be effectively tapped by Japan.

Early successes. The first years of the war brought Japan great success. In the Philippines, Japanese troops occupied Manila in January 1942, although Corregidor held out until May; Singapore fell in February, and the Dutch East Indies and Rangoon (Burma) in early March. The Allies had difficulty maintaining communications with Australia, and British naval losses promised the Japanese navy further freedom of action. Tōjō grew in confidence and popularity and began to style himself somewhat in the manner of a fascist leader. But the U.S. Navy had not been permanently driven from the South Pacific. The Battle of Midway in June 1942 cost the Japanese fleet four aircraft carriers and many seasoned pilots, and the battle for Guadalcanal Island in the Solomons ended with Japanese withdrawal in February 1943.

Japan on the defensive. After Midway, Japanese naval leaders secretly concluded that Japan's outlook for victory was poor. When the fall of Saipan in July 1944 brought U.S. bombers within range of Tokyo, the Tōjō cabinet was replaced by that of Koiso Kuniaki. Koiso formed a supreme war-direction council designed to link the cabinet and the high command. Many in government realized that the war was lost, but none had a program for ending the war that was acceptable to the military. There were also grave problems in breaking the news to the Japanese people, who had been told only of victories. Great firebombing raids in 1945 brought destruction to every major city except the old capital of Kyōto; but the generals were bent on continuing the war, confident that a major victory or protracted battle would help gain honourable terms. The Allied talk of unconditional surrender provided a good excuse to continue the fight.

In February 1945 the emperor met with a group of senior statesmen to discuss steps that might be taken. When U.S. landings were made on Okinawa in April, the Koiso government fell. The problem of the new premier, Admiral Suzuki Kantarō, was not whether to end the war but how best to do it. The first plan advanced was to ask the Soviet Union, which was still at peace with Japan, to intercede with the Allies. The Soviet government had agreed, however, to enter the war; consequently, its reply was delayed while Soviet leaders participated in the Potsdam Conference in July. The Potsdam Declaration issued on July 26 offered the first ray of hope with its statement that Japan would not be "enslaved as a race, nor destroyed as a nation."

Atomic bombings

The end of the war. Atomic bombs largely destroyed the cities of Hiroshima and Nagasaki on August 6 and 9, respectively. On August 8 the Soviet Union declared war and the next day marched into Manchuria, where the Kwantung Army could offer only token resistance. The Japanese government attempted to gain as its sole condition for surrender a qualification for the preservation of the imperial institution; after the Allies agreed to respect the will of the Japanese people, the emperor insisted on surrender. The Pacific war came to an end on August 14

(August 15 in Japan). The formal surrender was signed on September 2 in Tokyo Bay aboard the battleship USS *Missouri*.

Military extremists attempted unsuccessfully to prevent the radio broadcast of the emperor's announcement to the nation. There were a number of suicides among the military officers and nationalists who felt themselves dishonoured, but the emperor's prestige and personal will, once expressed, sufficed to bring an orderly transition. To increase the appearance of direct rule, the Suzuki cabinet was replaced by that of Prince Higashikuni Naruhiko.

Postwar investigators concluded that neither the atomic bombs nor the Soviet entry into the war was central to the decision to surrender, although they probably helped to advance the date. It was determined that submarine blockade of the Japanese islands had brought economic defeat by preventing exploitation of Japan's new colonies, sinking merchant tonnage, and convincing Japanese leaders of the hopelessness of the war. Bombing brought the consciousness of defeat to the people. The destruction of the Japanese navy and air force jeopardized the home islands. Japan's largest armies, however, were never defeated, and this was responsible for the army's eagerness to fight on. By the end of the war, Japan's cities were destroyed, its stockpiles exhausted, and its industrial capacity gutted. The government stood without prestige or respect. An alarming shortage of food and rising inflation threatened what remained of national strength.

THE EARLY POSTWAR DECADES

Occupation. From 1945 to 1952 Japan was under Allied military occupation, headed by the Supreme Commander for Allied Powers (SCAP), a position held by U.S. General Douglas MacArthur until 1951. Although nominally directed by a multinational Far Eastern Commission in Washington, D.C., and an Allied Council in Tokyo—which included the United States, the Soviet Union, China, and the Commonwealth countries—the occupation was almost entirely an American affair. While MacArthur developed a large General Headquarters in Tokyo to carry out occupation policy, supported by local "military government" teams, Japan, unlike Germany, was not governed directly by foreign troops. Instead, SCAP relied on the Japanese government and its organs, particularly the bureaucracy, to carry out its directives.

The occupation represented a period of rapid social and institutional change that was based on the borrowing and incorporation of foreign models. General principles for the proposed governance of Japan had been spelled out in the

U.S. policy

Potsdam Declaration and elucidated in U.S. government policy statements drawn up and forwarded to MacArthur in August 1945. The essence of these policies was simple and straightforward: the demilitarization of Japan, so that it would not again become a danger to peace; democratization, meaning that, while no particular form of government would be forced upon the Japanese, efforts would be made to develop a political system under which individual rights would be guaranteed and protected; and the establishment of an economy that could adequately support a peaceful and democratic Japan.

In the early months of the occupation, SCAP acted swiftly to remove the principal supports of the militarist state. All individuals prominent in wartime organizations and politics, including commissioned officers of the armed services and all high executives of the principal industrial firms, were removed from their positions. An international tribunal was established to conduct war crimes trials, and seven men, including the wartime prime minister Tōjō, were convicted and hanged; another 16 were sentenced to life imprisonment.

Political reform. The most important reform carried out by the occupation was the establishment of a new constitution. When Japanese efforts to write a new document proved inadequate, MacArthur's government section prepared its own draft and presented it to the Japanese government as a basis for further deliberations. Endorsed by the emperor, this document was placed before the first postwar Diet in April 1946. It was formally promulgated on November 3 and went into effect on May 3, 1947.

The emphasis in the new constitution was clearly on the people rather than the throne. Sovereignty now lay with the people. A 31-article bill of rights followed, with Article 9 renouncing forever "war as a sovereign right of the nation." The emperor, no longer "sacred" or "inviolable," was now described as the "symbol of the state and of the unity of the people." The constitution called for a bicameral Diet, with the greatest power concentrated in the House of Representatives, members of which would now be elected by both men and women. The old peerage was dissolved, and the House of Peers was replaced by a House of Councillors. The Privy Council was abolished. The prime minister was to be chosen by the Diet from its members, and an independent judiciary was established with the right of judicial review.

Despite its hasty preparation and foreign inspiration, the new constitution gained wide public support. Although the ruling conservatives desired to revise it after Japan regained its sovereignty in 1952, and an official commission favoured changes in the constitution in 1964, no political group in postwar Japan has been able to secure the two-thirds majority needed to make revisions. While parts of the structure established by the document have been modified through administrative actions—including a reversal of the principle of decentralization in areas such as the police, the school system, and some spheres of local administration—and while Article 9 has been compromised by the decision to form a National Police Reserve that in 1954 became the Self-Defense Forces, the basic principles of the constitution have enjoyed support among all factions in Japanese politics. Executive leadership proved to be the chief asset of the new institutions, and, with the abolition of the competing forces that had hampered the premiers of the 1930s, Japan's postwar prime ministers have found themselves firmly in charge of the administration and (with limited rearmament) the armed forces as well.

Executive leadership

Economic and social changes. SCAP was aware that political democracy in Japan required not only a weakening of the value structure of the hierarchic "family state," which restricted the individual, but also a liberation of the Japanese people from the economic forces that reinforced such a state. With nearly half of Japan's farmers subsisting as tenants, Americans saw little hope for democracy in Japan without significant changes in the ownership of land. Occupation authorities therefore set out to establish a program of land reform that was designed to convert tenants into owner-farmers. Through legislation a plan was devised whereby landlords, many of whom lived in the cities, were forced to divest themselves of a high proportion of their holdings to the government. This land was then sold to tenants on favourable terms. The reforms were implemented with great efficiency and in the end proved highly successful. Supported by favourable tax and price arrangements, the majority of Japan's new owner-farmers gained control of their land. Benefited by agricultural subsidies and government-maintained high agricultural prices, the Japanese countryside experienced increased prosperity. Rural voters became not only the mainstay of the conservative Liberal-Democratic Party (LDP) after its formation in 1955, but as one of Japan's most powerful lobbies they often successfully resisted agricultural trade liberalization. In a reversal of the Taishō dilemma that sprang from low domestic consumption, land reform and agricultural price supports contributed significantly to Japan's emergence as a consumer economy in the 1950s and '60s.

The dissolution of Japan's great financial houses (*zaibatsu*) also was an early occupation priority, but it gave way under Cold War pressures. Although the *zaibatsu* originally were seen as the chief potential war makers, the need for an economically viable Japan changed this perspective to viewing them as essential for economic recovery. New legislation sought to enforce fair trading and to guard against a return to monopolies. The war itself, new postwar tax policies, and the purges that removed many top executives further undercut the largest firms. By 1950 extensive changes, although far short of those initially proposed, had taken place in the industrial world. The large banks, however, were not broken up and proved to be the centres for a measure of reconsolidation in the years after the occupation ended.

Strengthening the influence of labour in Japan also was seen as important for the advancement of democracy. A new Ministry of Labour was established in 1947. Laws on trade unions and labour relations modeled on New Deal legislation in the United States were passed, and a strong union movement was initially encouraged. But a proposed general strike in 1947 and the Cold War-induced shift toward rapid economic reconstruction, anti-inflationary policies, and a control of radicalism quickly resulted in an effort to bring labour under government control. At the same time a new labour organization, the General Council of Trade Unions of Japan (Sōhyō), was sponsored as a counterweight and gradual replacement for the Congress of Industrial Labour Unions of Japan (Sambetsu Kaigi), which had become dominated by the left. In the late 1950s Sōhyō, too, had become increasingly antigovernment and anti-American, its Marxist and socialist orientation finding a political voice in the Japan Socialist Party (JSP), of which it became the leading supporter.

Sōhyō

Postwar social legislation also provided relief from earlier restrictions. The civil code, which had supported the power of the male family head in the past, was rewritten to allow for equality between the sexes and joint inheritance rights. Women were given the right to vote and to sit in the Diet.

Educational reforms. Occupation authorities, convinced that democracy and equality were best inculcated through education, revised the Japanese educational system. A Fundamental Law of Education was passed in 1947, which guaranteed academic freedom, extended the length of compulsory education from six to nine years, and provided for coeducation. The American 6-3-3-4 structure of elementary, lower secondary, higher secondary, and undergraduate higher education was adopted. Entrance to high schools and universities came to depend on passing highly competitive examinations. To complement Japan's prewar elite institutions, the Americans sought to encourage the establishment of prefectural universities and junior colleges.

Political trends. Politics under the occupation and new constitution experienced considerable flux, as many of Japan's prewar leaders found themselves purged from public office and the two prewar parties, the Seiyūkai and Minseitō, restructured themselves as the Liberal and Progressive parties, respectively (the latter eventually becoming the Japan Democratic Party). On the left wing, the socialists and communists also reorganized their respective parties. Initial postwar elections included many political splinter groups. Faced with a lack of consensus, cabinets tended to be unstable and short-lived. This was true of the first Yoshida Shigeru cabinet (1946–47), which implemented most of the early SCAP reforms. Yoshida's return to power in the fall of 1948 resulted in a more stable situation and ushered in the Yoshida era, which lasted until 1954. During those years, Japan capitalized on the economic benefits of close cooperation with the United States during the Korean War (1950–53), which laid the groundwork for national reconstruction and for the essential postwar U.S.-Japan relationship. In 1951 Yoshida achieved what he regarded as his greatest accomplishment—the restoration of national sovereignty—by taking Japan to the San Francisco peace conference. There, with the American negotiator John Foster Dulles and representatives of 47 nations, he hammered out the final details of the Treaty of Peace with Japan. The treaty was formally signed on Sept. 8, 1951, and the occupation of Japan ended on April 28, 1952.

The era of rapid growth. From 1952 to 1973 Japan experienced accelerated economic growth and social change. By 1952 Japan had at last regained its prewar industrial output. Thereafter, the economy expanded at unprecedented rates. At the same time, economic development and industrialization supported the emergence of a mass consumer society. Large numbers of Japanese who had previously resided in villages became urbanized; Tokyo, whose population stood at about three million in 1945, reached some nine million by 1970. The first trade frictions, over Japanese textile exports, took place at that time. Japan's new international image was projected and enhanced by events such as the highly successful 1964 Olympic Summer Games and the Ōsaka World Exposition of 1970.

Urbanization

Economic transformation. The Korean War marked the turn from economic depression to recovery for Japan. As the staging area for the United Nations forces on the Korean peninsula, Japan profited indirectly from the war, as valuable procurement orders for goods and services were assigned to Japanese suppliers. The Japanese economy at the return of independence in 1952 was in the process of growth and change. Sustained prosperity and high annual growth rates, which averaged 10 percent in 1955–60 and later climbed to more than 13 percent, changed all sectors of Japanese life. The countryside, where farmers had benefited from land reform, began to feel the effects of small-scale mechanization and a continuous migration to industrial centres. Agricultural yields rose as improved strains of crops and modern technology were introduced, as household appliances appeared in remote villages, and as the changing patterns of urban food consumption provided an expanded market for cash crops, fruits and vegetables, and meat products. Efforts to control population growth, which had begun with the legalization of abortion in 1948 and included a national campaign to encourage family planning, showed considerable success, as the population stabilized and thereafter grew slowly. Gains in economic output, therefore, were not offset by a rapidly expanding population, and steady industrial growth brought full employment and even labour shortages.

Two elements underscored rapid growth in the 1960s. The first was the development of a consumer economy, which was given a significant boost by Ikeda Hayato's Income Doubling Plan of 1960. The second was the new industrial policy that emerged out of the Ministry of Trade and Industry (MITI) in 1959. Under these influences the structure of the Japanese economy changed to concentrate on high-quality and high-technology products designed for domestic and foreign consumption. The production of such products also emphasized Japan's need for stable, economically advanced trading partners to replace the Asian markets to which inexpensive textiles had been sent earlier. Improvements in transportation—*e.g.,* cargo-handling methods and bulk transport by large ore carriers and tankers—helped to remove the disadvantage of the greater distances over which Japan's products had to be shipped. Most important, the large and growing domestic market was rendering invalid earlier generalizations about Japan's need for cheap labour and captive Asian colonies to sustain its economy. The era of high growth continued until the "oil shock" of 1973: the embargo by the Organization of Petroleum Exporting Nations (OPEC). In the interim, Japan's industrial expansion made it a world leader in shipbuilding, electronics, precision optical equipment, steel, automobiles, and high technology. In the 1960s Japanese exports expanded at an annual rate of more than 15 percent, and in 1965 Japan revealed the first signs that it had a trade surplus.

Trade surplus

A number of factors greatly aided Japan's economic resurgence during the 1950s and '60s. One was the complete destruction of the nation's industrial base by the war. This meant that Japan's new factories, using the latest developments in technology, were often more efficient than those of their foreign competitors. The Japanese became enthusiastic followers of the American statistician W. Edward Deming's ideas on quality control and soon began producing goods that were more reliable and contained fewer flaws than those of the United States and western Europe. At the same time, Japan was able to import, under license, advanced foreign technology at relatively low cost. With the addition of a youthful and well-educated workforce, a high domestic savings rate that provided ample capital, and an activist government and bureaucracy that provided guidance, support, and subsidies, the ingredients were in place for rapid and sustained economic growth.

Social change. Two major changes were visible in the social life of the Japanese from 1952 to 1973. The first was the significant decline in the birth rate that stabilized the Japanese population. The second was the population shift from the countryside to urban centres. But even with a stable population Japan remained one of the world's most densely populated countries.

As population growth slowed and the economy expanded, Japan faced a labour shortage that drew workers from agriculture, as well as from small and medium enterprises, to the new large-scale industries of the cities. The resulting shift in Japan's population was dramatic. In the Meiji period the rural population of Japan stood at 85 percent of the national total; by 1945 it was approximately 50 percent, and by 1970 it had fallen to less than 20 percent. Agriculture itself became increasingly mechanized and commercialized. As sons, and even husbands, went off to the factories, women, children, and the elderly were often left to run the family farm. By 1970 the average farm household income had risen higher than its urban counterpart, providing considerable rural purchasing power. Television tied rural households to urban Japan and to the world beyond. Young men brought up on visions of urban life as projected by American television programs were eager to move to the cities after graduation from high school. Young women showed increasing reluctance to become farm wives. Rural solidarity suffered from such out-migration, and in many cases prewar village life ceased to be, as villages amalgamated into cities and struggled to develop new identities.

Cities also underwent rapid change. By 1972 one in every nine Japanese lived in Tokyo and one in four lived in the Tokyo-Ōsaka industrial corridor. As the national centre for government, finance, business, industry, education, and the arts, Tokyo became a magnet for many Japanese and the quintessential expression of Japanese urban life.

But while Tokyo and other large cities remained highly attractive, urban dwellers also faced serious problems, notably housing. Living space for most urban dwellers was infinitesimal when compared with Western societies. The dream of owning one's home, which most urban dwellers sought to keep alive, was already becoming increasingly elusive by the 1970s. In 1972 the price of land in or near Japan's largest cities was some 25 times higher than it had been in 1955, far surpassing the rise in the average urban worker's disposable income for the same period. While government and private industry were able to provide some low-cost housing, higher-priced housing in the form of high-rise condominiums, or "mansions," proliferated, and for most Japanese urbanites housing remained the chief flaw in Japan's postwar economic "miracle."

High land prices

Urban life also brought about changes in traditional Japanese family and gender relationships. The position of women improved, as many more now went to high schools and colleges. Most found urban employment until marriage. As arranged marriages declined and "love" matches increased, marriage customs also changed. Urban living promoted the ideal of the nuclear family, particularly as housing conditions made it difficult for the extended family to live together. Urban dwellers found themselves less dependent on the goodwill of their neighbours. There was also less need for the conformity that typified rural life—although for many recent arrivals the city-based company and factory effectively restructured village values to support an efficient workplace.

The majority of villagers actually made the transition from rural to urban life with less social stress than was the case in Europe and America. Juvenile delinquency showed some increase, but overall crime rates remained low. So-called "new" religions such as Sōka Gakkai (Value-Creation Society), which strongly appealed to those feeling isolated or alienated, flourished in the 1950s and '60s. Disparities between the newly rich and the older generation living on fixed incomes and between a freer, franker, and often more egotistic and brash mass culture that appealed to the young and traditional taste set by what once had been the aristocracy often accentuated how generations viewed the postwar situation.

Political developments. With the restoration of sovereignty, politicians who had been purged by the occupation were allowed to return to public life. This included a number of prewar rightists who had been active in the 1930s. But the ideological right found few adherents among the postwar generation, and without military or big-business support the right wing played a largely dormant role during the 1950s and '60s. The left fared considerably better. Communists who returned to Japan from foreign exile or who were released from domestic prisons played a vigorous role in the immediate postwar political arena. In 1949 the Japan Communist Party (JCP) elected 35 candidates to the lower house and garnered 10 percent of the vote. But by 1952 the Korean War (which had led SCAP to purge communists from public office), steady improvements in living conditions, and uncooperative Soviet attitudes in negotiations over the return of the Kuril Islands and over fishing treaties had seriously undermined public support for the communists.

The year 1955 was highly significant in postwar politics. The right and left wings of the socialist movement, which had been divided since 1951 over the peace treaty, merged to form the Japan Socialist Party (JSP). Faced with this united opposition the conservative parties, the Liberals and the Democrats, joined to found the Liberal-Democratic Party (LDP). Japan thus entered a period of essentially two-party politics. The dominant LDP, which inherited Yoshida's mantle, worked effectively to solidify the close ties he had created with bureaucrats, bankers, and the business community. As a result, ex-bureaucrats played significant roles in the LDP, often being elected to the Diet and becoming important cabinet members. Three of the next six prime ministers (all from the LDP) who succeeded Yoshida—Kishi Nobusuke, Ikeda Hayato, and Satō Eisaku—were ex-bureaucrats.

Liberal-Democratic Party

Ideologically, the LDP combined a strong commitment to economic growth with the desire to return Japan to world prominence. The party depended on the financial support of business and banking, but its voter base remained in rural Japan. At the local level, LDP politicians established political networks that became the hallmarks of postwar politics and emphasized the role of personal "machine" politics over party platforms. But individual LDP Diet members realized that in order to provide patronage for their constituents they needed the support of party leaders with access to the bureaucracy. Factions therefore formed around such leaders, who vied with one another for the premiership and sought to have members of their faction appointed to important cabinet posts.

As the voice of the opposition, the JSP resisted rearmament, had a strong antinuclear stance, campaigned to rid Japan of the American bases and abrogate the Mutual Security Treaty, supported mainland China, and vigorously opposed all efforts to change the postwar constitution. The appeal of the JSP was directed both to urban intellectuals and to the working classes, and its financial support came largely from labour (Sōhyō). In contrast to the LDP's focus on economic growth, big business, and agriculture, the JSP concentrated on urban issues, on those bypassed by prosperity, and on the mounting problems of pollution and environmental degradation that accompanied accelerated industrial growth. Socialist influence was weakened, however, when the more right-wing JSP members split off to form the Democratic Socialist Party (DSP) in 1959.

By the early 1970s urban issues also attracted the JCP, which started to substitute practical matters for ideology and won a number of mayoral elections. To the right of the communists and socialists appeared the Clean Government Party (Kōmeitō), which began in 1964 as the political arm of Sōka Gakkai but dissociated itself from the religion by 1970; like its opposition counterparts, it focused on the urban electorate. On occasion, as in 1960 with the Kishi government and the proposed renewal of the U.S-Japan Mutual Security Treaty, the opposition could mount sufficient public support to bring down an LDP cabinet, but on the whole the era was one in which the LDP remained firmly in power.

Still, by the late 1960s and early '70s there also were signs of a decline in LDP support. In 1952 the LDP had garnered two-thirds of the Diet seats, but by 1972 it controlled only slightly more than half. The effects of the so-called "Nixon shocks" in 1971, which allowed the yen to rise against the dollar and restructured the U.S.-China (and

The "Nixon shocks"

hence the Japan-China) relationship, were compounded in 1973 by the OPEC oil crisis that threatened the underpinnings of Japan's postwar prosperity and the LDP's political hegemony.

International relations. The Japan that returned to the international community in 1952 was considerably reduced in territory and influence. The Republic of China (Taiwan), the People's Republic of China on the mainland, the Republic of Korea (South Korea), and the Democratic People's Republic of Korea (North Korea) all possessed military establishments far larger than what became Japan's Self-Defense Forces. The United States maintained its occupancy of Okinawa and the Ryukyus, while the Soviet Union occupied the entire Kuril chain and claimed southern Sakhalin. The Korean War increased the urgency for a peace treaty.

The San Francisco peace conference that convened in September 1951 ratified arrangements that had been worked out earlier. In the peace treaty that ensued, Japan recognized the independence of Korea and renounced all rights to Taiwan, the Pescadores, the Kurils, and southern Sakhalin and gave up the rights to the Pacific islands earlier mandated to it by the League of Nations. The Soviet Union attended the conference but refused to sign the treaty. This enabled Japan to retain hope for regaining four islands of the Kurils closest to Hokkaido—territory that Japan had gained through negotiations, not war. The peace treaty recognized Japan's "right to individual and collective self-defense," which it exercised through the United States–Japan Security Treaty (1951) by which U.S. forces remained in Japan until the Japanese secured their own defense. Japan agreed not to grant similar rights to a third power without U.S. approval. Americans promised to assist Japan's Self-Defense Forces while U.S. military units (except air detachments and naval forces) were withdrawn to Okinawa.

The treaty made no arrangements for reparations to the victims of Japan's Pacific war but provided that Japan should negotiate with the countries concerned. Consequently, effective resumption of relations with the nations of Asia came only after treaties covering reparations had been worked out with them. These were signed with Burma (now Myanmar) in 1954, the Philippines in 1956, and Indonesia in 1958. In 1956 Japan restored diplomatic relations with the Soviet Union but without a formal peace treaty. With the Soviet Union no longer blocking the way, Japan was admitted to the United Nations in late 1956 and subsequently became active in UN meetings and specialized agencies. Japan spearheaded the creation of the Asian Development Bank in 1965–66.

At the time of the peace treaty, Prime Minister Yoshida wanted to delay committing Japan to either of the two Chinas, but the U.S. negotiator John Foster Dulles convinced him that the treaty would be opposed in the U.S. Senate unless assurances were given that Japan would recognize the Republic of China. Thus, Tokyo soon negotiated a peace treaty with that regime, but one that would not prejudice subsequent negotiations with Peking. A lively trade developed with Taiwan, where Japan made considerable contributions to the economy.

Trade with China

Trade relationships with mainland China developed slowly in the absence of diplomatic ties. In 1953 an unofficial trade pact was signed between private Japanese groups and Chinese authorities. In addition, a semiofficial "memorandum" trade became increasingly important in the 1960s. The Chinese government made skillful use of trade for political purposes. Nevertheless, Japan gradually became China's most important trading partner.

U.S. overtures toward mainland China in 1971 led to a rapid reorientation of Japan's China policy. Japanese government leaders indicated a willingness to compromise ties with Taiwan in favour of a closer relationship with Peking. Peking also revealed a new interest in formal relations with Japan, subject to Japan's revocation of its treaty with Taiwan. In 1972, a year after mainland China was admitted to the UN, Prime Minister Tanaka reached an agreement with Peking on steps to normalize relations. Japan simultaneously severed its ties with Taiwan, replacing its embassy with a nonofficial office.

Japan's post-occupation relationship with the United States was founded on the 1951 security treaty. Part of the understanding that lay behind this treaty was that Japan would have access to the U.S. market in exchange for the maintenance of American bases on Japanese soil. While the LDP saw advantages to maintaining such a quid pro quo relationship, which allowed Japan to dramatically expand its foreign trade while avoiding undue security costs, Japan's opposition parties were less sanguine about a relationship that tied Japan directly into the increasingly hostile Cold War. Tensions therefore mounted as the renewal date of the treaty (scheduled for 1960) approached; both governments hoped to extend it for 10 years as the revised Treaty of Mutual Cooperation and Security. When the Kishi cabinet used its majority in the Diet to force through treaty revisions, opposition increased steadily. Gigantic public demonstrations, largely composed of students, shook Tokyo for days. In the end the treaty survived, but Kishi resigned in July 1960.

The administration of U.S. President John F. Kennedy caught the imagination of many Japanese, and Kennedy's designation of the popular scholar Edwin O. Reischauer as ambassador further improved Japanese-American relations. But by the late 1960s the unpopularity of the Vietnam War threatened to disturb the relationship once more. Prime ministers Ikeda and Satō worked hard to remove the final reminders of war. In 1967, under Satō, the Bonin (Ogasawara) Islands were restored to Japan; and in 1969, on the eve of renewed negotiations over treaty revisions, the United States agreed to return the Ryukyus in 1972, although bases were to be maintained on Okinawa under the terms of the security treaty. The treaty was renewed without incident in 1970, now changed to allow termination by either side with a year's advanced notification. Thus, by 1972 the U.S.-Japan relationship had stabilized and there was as yet little to indicate the mounting conflict over trade that subsequently emerged. (M.B.J./F.G.N.)

JAPAN SINCE 1973

End of high growth

Economic change. By the early 1970s a series of forces ended the of high growth that Japan had experienced since the 1950s. These included advances in technology, the disappearance of rural labour for industry, and the decline in international competitiveness of such heavy industries as shipbuilding, aluminum, fertilizers, and steel. Outcries over urban congestion, pollution, and environmental degradation and dissatisfaction with escalating land prices caused many middle-class Japanese to question the logic that linked growth with national success. The international environment also was changing. In 1971 the United States devalued the U.S. dollar by 17 percent against the Japanese yen. In 1973–74 the OPEC oil embargo created a further disruption of the Japanese economy, which depended heavily on Middle Eastern oil. Japanese dependency on fuel and food—the United States had temporarily embargoed soybean exports to Japan in 1972—had become increasingly clear.

During the 1970s and '80s, consequently, Japan tried to adjust its economy to changing global conditions and sought to diversify its markets and sources of raw materials. Japan called for free trade and tried to create some energy self-sufficiency through nuclear power generation. The economic uncertainties of the 1970s revived a defensive, nationalistic sentiment that pictured Japan in a struggle with outside forces aimed at depriving the Japanese of their hard-won postwar gains. Until the early 1990s, international economic tensions were used by the ruling LDP and the bureaucracy to contain and defuse domestic economic and political issues.

The domestic rhetoric about the hostile international environment Japan faced diverted attention from the fact that by the 1980s the Japanese economy had become one of the world's largest and most sophisticated. Per capita income had surpassed that of the United States, and total gross national product stood at roughly one-tenth of world output. By the mid-1980s Japan had become the world's leading net creditor nation and the largest donor of development aid. Prosperity, however, increasingly depended on trade. Slow domestic growth was offset by booming exports. In the 1970s exports were seen as a counterbalance to the deficits anticipated from rapidly rising oil prices. As the Japanese economy weathered the recessions induced by escalating oil prices in 1972–74 and 1979–81, exports grew. Headed by automobiles, colour television sets, high-quality steel, precision optical equipment, and electronic products, Japan's merchandise trade surplus with western Europe and the United States steadily mounted.

These surpluses increased friction between Japan and its trading partners in Europe and the United States. Accusations were made that Japan advocated free trade abroad but maintained a closed market at home, that Japan engaged in "adversarial trade" designed to benefit only Japan, that Japan exported domestic unemployment during economic hard times, and that Japan sold goods at lower prices abroad than at home. The government and bureaucracy responded by making efforts to "open" Japan. By the end of the 1980s most formal trade barriers had been removed, although various types of informal nontariff barriers remained.

The "bubble economy"

Beginning in the mid-1980s there emerged in Japan what came to be called the "bubble economy," characterized by easy credit and unbridled speculation that drove Japanese equity and real estate markets to astronomical price levels. The bursting of this bubble in the early 1990s ushered in Japan's worst postwar recession, the severity and duration of which have postponed planned reforms, undermined consumer confidence, and exacerbated trade tensions. Japan's merchandise trade surplus with the world continued to spiral, leading in the mid-1990s to an appreciation of the yen against the dollar.

Well before the bubble economy and its demise, the Japanese economy had already showed clear signs of maturation. Growth rates were modest, the average age of the workforce was rising rapidly, and worker productivity—particularly in the white-collar sector, which had grown to roughly half the workforce—was declining and inhibiting profitability. While unemployment remained at low levels by Western standards, layoffs became more common. In many industries salaries became tied to performance rather than seniority, and the traditional policies of lifetime employment and company paternalism began to disappear. Faced with domestic labour shortages and high wages, Japanese industrialists turned increasingly to offshore production to remain competitive. Since the early 1990s, however, the recession has been the defining characteristic of Japan's economy.

Political developments. The LDP continued to dominate Japanese politics until 1993. The success of the party in steering Japan through the difficult years of the OPEC oil crisis and the economic transition that substituted high-technology enterprises for smokestack industries in the 1970s and '80s, thereby restoring Japan's international economic confidence, was not lost on the Japanese people. By the late 1980s and early '90s, however, as economic growth slowed and income disparities heightened public sensitivity to political corruption, this bargain between the people and their government changed.

Yet, there were also earlier signs of a political transition. While LDP rule appeared to be strengthening, the party's share of the popular vote was declining. And, while the premiership remained firmly under LDP control, all governments but that of Nakasone Yasuhirō (1982–87) were short-lived. In 1989 the LDP lost control of the upper house to a coalition of opposition parties headed by the socialists.

The era had begun in 1972 with considerable hope for political change, as Tanaka Kakuei, a self-made politician who defied the usual LDP bureaucratic model, sought to address the problems of pollution and urban crowding by calling for a redistribution of industry throughout the Japanese islands. Tanaka's era ended in 1974 with little change and with Tanaka mired in a major influence-peddaling scandal. Indeed, Tanaka came to symbolize the rise of "money politics," as election campaigns became increasingly expensive and faction leaders—expected to provide campaign funds to their followers—became heavily entangled in questionable financial relationships. Although there were early calls for reform, few in the LDP were prepared to make changes. To some degree Tanaka, who was arrested in 1976 and convicted of bribery charges in 1983, underscored this reluctance on the part of the LDP to undertake serious reforms. By the late 1980s, political corruption had become almost endemic, and the LDP was racked by a succession of scandals.

Death of Hirohito

Political turmoil was muted for some months during Emperor Hirohito's illness in 1988. His death, in January 1989, ended the Shōwa era, the longest recorded reign in Japanese history—some 62 years. He was succeeded by his son, Akihito, who took the reign name Heisei ("Achieving Peace").

Later in 1989 Prime Minister Takeshita Noboru was forced out of office for involvement in an even bigger postwar scandal involving manipulation of the stock market. Takeshita's successor Uno Sōsuke almost instantly found himself in a sex scandal. Uno was replaced by the "clean" Kaifu Toshiki, who lacked firm support in the party. This became apparent during the Persian Gulf Crisis in 1991, when Kaifu found himself labeled "reluctant" and "indecisive" in handling Japan's response to American requests for assistance.

Miyazawa Kiichi, who succeeded Kaifu in 1991, tried to introduce reform legislation in the Diet. This cost him the support of key LDP members, and a no-confidence motion in June 1993, with many LDP members participating, brought down his government. In elections held the next month, the LDP lost its Diet majority to a coalition of opposition parties, ending its 38-year rule.

The July 1993 election ushered in an era of political transition. Several new parties emerged that were essentially splinter groups off the LDP, including the Japan New Party (JNP) and the Japan Renewal Party. These joined several former opposition parties to form a coalition government with Hosokawa Morihirō, leader of the JNP, as prime minister.

Hosokawa achieved some success in limiting contributions and managed to pass a modified elections package that included the creation of 300 single-member constituencies. Opposition within his coalition to tax reform and accusations of his own involvement in the Miyazawa-era scandal forced his resignation in April 1994. Hosokawa's successor lasted a mere two months. In the ensuing power vacuum, socialists and remaining LDP members formed an unlikely coalition, and Murayama Tomiichi became Japan's first socialist premier since 1948. Since then, the LDP has managed to stay in power through a series of shifting coalitions with smaller parties. At the beginning of the 21st century the major parties were the LDP, Democratic Party of Japan, New Komeito, Liberal Party, Japan Communist Party, Social Democratic Party of Japan, and the New Conservative Party.

Social change. Despite the long recession, Japan has continued its transformation into a high-technology, urban, industrial society. The migration from countryside to city largely has been completed; fewer than 10 percent of Japanese families now live on a farm. Urbanization has resulted in further demographic change, including an accelerating decline in the birth rate that by the mid-1980s was less than the level needed to replace the population. Thus, Japanese society faces a rapidly aging population and concomitant declining active workforce.

Japanese values also have been changing as generations born and raised in the city mature and replace those brought up in the villages. While Japanese society remains formally hierarchical and social distinctions based on education and family background persist, the degree of conformity and the acceptance of consensus appear to be lessening. Such phrases as "my car," "my home," and "my leisure" further underscore the growing emphasis on the individual and individual choice and on the more assertive attitude of the ordinary Japanese.

Gender relations

Gender relations also have undergone a gradual transition. Women now account for about two-fifths of the workforce, but most occupy temporary or part-time positions. Women, particularly those married to white-collar workers, are still expected to carry much of the responsibility of household management and child rearing, while the males devote themselves to their office culture.

"Internationalization" has been another important theme since the early 1970s, as large numbers of Japanese have traveled abroad and an increasing number of foreign students and foreign workers have come to Japan. Foreign labourers from the Middle East and South and Southeast Asia now perform many of the less desirable jobs. The absorption of such residents has not always been easy for a society that sees itself as racially distinct and homogeneous. Discrimination against minorities, however—including Koreans, the former outcast group now called *burakumin,* and the Ainu—which has persisted for centuries, appears less acceptable today in a society that is increasingly subject to international scrutiny and criticism. The internationalization of Japan also has resulted in a reassertion of Japanese nationalism, particularly among the older members of society who see Japan losing its identity amid the influx of foreign culture. The sense of national purpose has also been affected by the long recession.

International relations. Japan has continued to pursue a policy of close cooperation with the United States since the early 1970s while rebuilding relations with its Asian neighbours. Despite the rapid political transformation of the world after the end of the Cold War, ties between the United States and Japan have been little altered in their fundamental tenets. Both nations remain committed to the Mutual Security Treaty, which keeps Japan under the U.S. nuclear weapons "umbrella" and permits thousands of American troops to be stationed there. Not to be overlooked are the growing tensions between Japan and the United States that have emerged with the maturation of the Japanese economy and Japan's growing international involvement. Clearly the power relationship between them has been shifting from one akin to teacher and student to that of colleagues. Trade issues have been particularly acrimonious. As the recession has persisted, trade disputes have been increasingly overshadowed by bilateral discussions of how Japan can revive its economy.

Japan has also been criticized, especially by European countries, for viewing its national interest and foreign policy in narrow economic terms and for its unwillingness to respond to international problems and crises in a manner commensurate with its economic stature. Following Prime Minister Tanaka's trip to China in 1972, which began the process of normalizing relations between the two nations, Japan vigorously pursued trade opportunities with China, and in 1978 the two countries concluded a peace treaty. Both trade and cultural contacts between Japan and China have expanded dramatically since then, and in the early 1990s China became Japan's second largest trading partner, surpassed only by the United States. The visit to China by Emperor Akihito in 1992, which included a tacit apology for the "severe suffering" that the Japanese had inflicted on the Chinese during the war, along with similar statements made by various prime ministers since then, showed that Japan was determined not just to build economic ties with China but to transcend the gap that stemmed from the war and to restore cultural ties.

Although Japan's formal relationship with Taiwan was discontinued after 1978, Taiwan has continued to play an important role for Japan, particularly since 1987 when Japan began to deepen its ties with the so-called newly industrialized countries of Asia: South Korea, Taiwan, Hong Kong, and Singapore.

Efforts to solidify relations with Southeast Asia have also moved ahead since 1973. Southeast Asian nations—particularly Indonesia—became recipients of extensive Japanese development aid. Japan also has made efforts to work with Vietnam and Cambodia. Japan's interests in Vietnam have been largely economic, but in Cambodia unarmed troops from Japan's Self-Defense Forces participated in a UN peacekeeping operation, the first time since the war that Japanese forces had ventured overseas.

The "northern territories"

Relations with Russia, however, have remained decidedly cool. A formal peace treaty was not worked out with the Soviet Union before its dissolution. The major sticking point on the Japanese side has been the disposition of the "northern territories," the four small islands in the southern Kuril chain that the Russians seized following the war. The Japanese have sought the return of these islands and have been reluctant to grant Russia development aid without an agreement.

The period since 1973 has set the groundwork for a major transformation of Japan's international role. With the maturation of its economy, Japan has entered the circle of the advanced industrial nations. Japan has cautiously played a more active role in international institutions. While many Japanese are reluctant for Japan to pursue an "activist" foreign policy role—particularly one hinting at military participation—Japanese leaders are aware of increasing international pressure to contribute more than just money to foreign policy. Conversely, many Japanese argue that full participation in world affairs should include a voice in decision making—specifically, a permanent seat on the UN Security Council and, with it, the long-sought legitimation of its status as a world leader. (M.B.J./F.G.N./Ed.)

For later developments in the history of Japan, see the BRITANNICA BOOK OF THE YEAR.

For coverage of related topics in the *Macropædia* and *Micropædia,* see the *Propædia,* sections 934, 96/10, and 975, and the *Index.*

BIBLIOGRAPHY

General works. Two excellent compendiums are RICHARD BOWRING and PETER KORNICKI (eds.), *The Cambridge Encyclopedia of Japan* (1993); and *Kodansha Encyclopedia of Japan,* 9 vol. (1983), and supplement (1986). Also useful are *Japan: An Illustrated Encyclopedia,* 2 vol. (1993); RONALD E. DOLAN and ROBERT L. WORDEN (eds.), *Japan: A Country Study,* 5th ed. (1992); and *Japan: Profile of a Nation* (1995). Further research can be found in *Bibliography of Asian Studies* (annual).

Representative atlases of Japan include TEIKOKU-SHOIN CO. (ed.), *Atlas Japan: In English & Japanese* (1989); MARTIN COLLCUTT, MARIUS JANSEN, and ISAO KUMAKURA, *Cultural Atlas of Japan* (1988); AKIRA EBATO and KAZUO WATANABE, *Atlas of Japan: Physical, Economic, and Social,* 2nd rev. ed. (1974); *Grand Atlas of Japan* (1985, reissued 1990), published by Heibonsha; and GEOGRAPHICAL SURVEY INSTITUTE, *The National Atlas of Japan,* rev. ed. (1990).

Physical and human geography. *The land:* RYUZIRO ISIDA, *Geography of Japan* (1961), is a comprehensive study. GLENN THOMAS TREWARTHA, *Japan* (1965, reissued 1979), is considered a classic regional geography. ASSOCIATION OF JAPANESE GEOGRAPHERS (eds.), *Geography of Japan* (1980), is highly regarded. Representative geographic works include *Regional Geography of Japan,* trans. from Japanese, 6 vol. (1957); NORTON GINSBURG, "Economic and Cultural Geography," in ARTHUR E. TIEDEMANN (ed.), *An Introduction to Japanese Civilization* (1974), pp. 423–459; GIL LATZ, "The Experience of Place in Japan," *Asian Art,* 5(2):2–7 (Spring 1992); and JACQUES PEZEU-MASSABUAU, *The Japanese Islands: A Physical and Social Geography* (1978; originally published in French, 1968). TOSHIO NOH and JOHN C. KIMURA (eds.), *Japan: A Regional Geography of an Island Nation,* 2nd ed. (1989), reviews the distinctive regional character of the archipelago. Geomorphology is covered by TORAO YOSHIKAWA, SOHEI KAIZUKA, and YOKO OTA, *The Landforms of Japan* (1981). Climatology is dealt with in E. FUKUI (ed.), *The Climate of Japan* (1977). A detailed survey of biogeography can be found in M. NUMATA (ed.), *The Flora and Vegetation of Japan* (1974). A classic field survey of village life is RICHARD K. BEARDSLEY, JOHN W. HALL, and ROBERT E. WARD, *Village Japan* (1959, reissued 1969). Japan's cities are treated in *Japanese Cities: A Geographical Approach* (1970); and DAVID KORNHAUSER, *Japan,* 2nd ed. (1982).

The people: Three penetrating profiles of Japanese society include CHIE NAKANE, *Japanese Society,* rev. ed. (1973, reissued 1984; originally published in Japanese, 1967); TAKEO DOI, *The Anatomy of Dependence* (1973, reissued 1981; originally published in Japanese, 1971); and TAKIE SUGIYAMA LEBRA (ed.), *Japanese Social Organization* (1992). GEORGE DE VOS and HIROSHI WAGATSUMA, *Japan's Invisible Race,* rev. ed. (1972), discusses the *burakumin* minority. Studies of Japanese religions include IAN READER, *Religion in Contemporary Japan* (1991); and H. NEILL MCFARLAND, *The Rush Hour of the Gods: A Study of New Religious Movements in Japan* (1967, reissued 1970). Gender questions are explored in TAKIE SUGIYAMA LEBRA, *Japanese Women: Constraint and Fulfillment* (1984); and SUMIKO IWAO, *The Japanese Woman* (1993).

The economy: General background texts include G.C. ALLEN, *A Short Economic History of Modern Japan,* 4th ed. (1981); and PENELOPE FRANCKS, *Japanese Economic Development* (1992). Energy and power issues are treated in RICHARD J. SAMUELS, *The Business of the Japanese State* (1987). Cogent analyses of the Japanese agricultural sector can be found in RONALD P. DORE, *Land Reform in Japan* (1959, reissued 1985), and *Shinohata* (1978, reissued 1994).

Other valuable works on the economy are TAKAFUSA NAKAMURA, *The Postwar Japanese Economy,* 2nd ed. (1995; origi-

nally published in Japanese, 1978); ARDATH W. BURKS, *Japan: A Postindustrial Power,* 3rd ed. rev. and updated (1991); FRANK GIBNEY, *Miracle By Design* (1982), and *Japan: The Fragile Superpower,* 2nd rev. ed. (1985); EDWARD J. LINCOLN, *Japan's New Global Role* (1993); KENNETH B. PYLE, *The Japanese Question: Power and Purpose in a New Era* (1992); and YOSHIO SUZUKI (ed.), *The Japanese Financial System* (1987, reissued 1992; originally published in Japanese, 1986).

Administration and social conditions: CHALMERS JOHNSON, *MITI and the Japanese Miracle* (1982), offers a seminal contribution to understanding Japan's government-business relationships. Thoughtful discussions of the Japanese political system include HANS H. BAERWALD, *Japan's Parliament* (1974); GERALD L. CURTIS, *The Japanese Way of Politics* (1988); TAKESHI ISHIDA and ELLIS S. KRAUSS (eds.), *Democracy in Japan* (1989); and DANIEL I. OKIMOTO and THOMAS P. ROHLEN (eds.), *Inside the Japanese System* (1988).

Reviews of the Japanese educational system can be found in WILLIAM K. CUMMINGS, *Education and Equality in Japan* (1980); THOMAS P. ROHLEN, *Japan's High Schools* (1983); and MERRY WHITE, *The Japanese Education Challenge* (1987).

Cultural life: H. PAUL VARLEY, *Japanese Culture,* 3rd ed. (1984), is the most useful cultural history. Cultural aspects also are treated in the encyclopaedias cited above. Additional sources may be found in the bibliographies to the articles EAST ASIAN ARTS and JAPANESE LITERATURE. (G.La.)

History. *General works:* The classic survey is GEORGE BAILEY SANSOM, *A History of Japan,* 3 vol. (1958–63, reissued 1978). The most comprehensive and detailed account in English is JOHN W. HALL *et al.* (eds.), *The Cambridge History of Japan* (1988–). Shorter, excellent interpretive works include EDWIN O. REISCHAUER, *Japan: The Story of a Nation,* 4th ed. (1990); EDWIN O. REISCHAUER and ALBERT M. CRAIG, *Japan: Tradition & Transformation,* rev. ed (1989); EDWIN O. REISCHAUER and MARIUS B. JANSEN, *The Japanese Today: Change and Continuity,* enlarged ed. (1995); and JOHN WHITNEY HALL, *Japan from Prehistory to Modern Times* (1970, reissued 1991). Works covering more specific periods include MIKISO HANE, *Premodern Japan* (1991), and *Modern Japan,* 2nd ed. (1992); CONRAD TOTMAN, *Japan Before Perry* (1981); and W.G. BEASLEY, *The Rise of Modern Japan* (1990).

Ancient Japan: Useful works on Japan's ancient history include C. MELVIN AIKENS and TAKAYASU HIGUCHI, *Prehistory of Japan* (1982); GINA L. BARNES, *Protohistoric Yamato* (1988); J.E. KIDDER, JR., *Japan Before Buddhism,* rev. ed. (1966); MICHIKO YAMAGUCHI AOKI (trans.), *Izumo Fudoki* (1971; originally published in Japanese, 1931); ROBERT S. ELLWOOD, *The Feast of Kingship* (1973); and RICHARD J. MILLER, *Ancient Japanese Nobility* (1974).

The Nara and Heian periods are discussed in ROBERT BORGEN, *Sugawara no Michizane and the Early Heian Court* (1986, reissued 1994); G. CAMERON HURST III, *Insei* (1976); IVAN MORRIS, *The World of The Shining Prince* (1964, reissued 1994); KARL F. FRIDAY, *Hired Swords* (1992); and WILLIAM WAYNE FARRIS, *Heavenly Warriors: The Evolution of Japan's Military, 500–1300* (1992).

Medieval Japan: Among the recent works of note are several by JEFFREY P. MASS: *Warrior Government in Early Medieval Japan* (1974, reissued 1991), *The Development of Kamakura Rule, 1180–1250* (1979), and *Lordship and Inheritance in Early Medieval Japan* (1989). Excellent volumes of essays by Japanese and Western scholars include JEFFREY P. MASS (ed.), *Court and Bakufu in Japan* (1982); JOHN WHITNEY HALL and JEFFREY P. MASS (eds.), *Medieval Japan* (1974, reissued 1988); and JOHN WHITNEY HALL and TOYODA TAKESHI (eds.), *Japan in the Muromachi Age* (1977). Other useful works include H. PAUL VARLEY, *Imperial Restoration in Medieval Japan* (1971); MINORU SHINODA, *The Founding of the Kamakura Shogunate, 1180–1185* (1960); KENNETH ALLEN GROSSBERG, *Japan's Renaissance* (1981); THOMAS KEIRSTEAD, *The Geography of Power in Medieval Japan* (1992); and HITOMI TONOMURA, *Community and Commerce in Late Medieval Japan* (1992).

Early modern Japan (1550–1850): CONRAD TOTMAN, *Early Modern Japan* (1993), is a good general text. Among the many excellent collections of essays are JOHN WHITNEY HALL and MARIUS B. JANSEN (eds.), *Studies in the Institutional History of Early Modern Japan* (1968); JOHN WHITNEY HALL, NAGAHARA KEIJI, and KOZO YĀMAMURA (eds.), *Japan Before Tokugawa* (1981; originally published in Japanese, 1978); CHIE NAKANE and SHINZABURŌ ŌISHI (eds.), *Tokugawa Japan* (1990); and TETSUO NAJITA and IRWIN SCHEINER (eds.), *Japanese Thought in the Tokugawa Period, 1600–1868* (1978, reissued 1988). The Sengoku era is discussed in MARY ELIZABETH BERRY, *The Culture of Civil War in Kyoto* (1994); GEORGE ELISON, *Deus Destroyed: The Image of Christianity in Early Modern Japan* (1973, reissued 1988); and GEORGE ELISON and BARDWELL L. SMITH (eds.), *Warlords, Artists & Commoners* (1981).

The political history of the Tokugawa period is best covered in CONRAD TOTMAN, *Politics in the Tokugawa Bakufu, 1600–1843* (1967), and *The Collapse of the Tokugawa Bakufu, 1862–1868* (1980); and HAROLD BOLITHO, *Treasures Among Men: The Fudai Daimyo in Tokugawa Japan* (1974). Economic history is discussed in THOMAS C. SMITH, *The Agrarian Origins of Modern Japan* (1959, reissued 1984); SUSAN B. HANLEY and KOZO YAMAMURA, *Economic and Demographic Change in Preindustrial Japan, 1600–1868* (1977); CHARLES DAVID SHELDON, *The Rise of the Merchant Class in Tokugawa Japan, 1600–1868* (1958, reprinted 1973); and TETSUO NAJITA, *Visions of Virtue in Tokugawa Japan* (1987). Excellent scholarship on Tokugawa thought is found in MASAO MARUYAMA, *Studies in the Intellectual History of Tokugawa Japan* (1974, reissued 1989; originally published in Japanese, 1965); and HERMAN OOMS, *Tokugawa Ideology* (1985). E.H. NORMAN, *Origins of the Modern Japanese State,* ed. by JOHN W. DOWER (1975), is a collection of representative essays; the themes are continued in two studies of peasant revolts: HERBERT P. BIX, *Peasant Protest in Japan, 1590–1884* (1986); and ANNE WALTHALL, *Social Protest and Popular Culture in Eighteenth-Century Japan* (1986). RONALD P. TOBY, *State and Diplomacy in Early Modern Japan* (1984, reissued 1991), discusses foreign relations.

Japan since 1850: Works on the opening of Japan and the fall of the shogunate include OLIVER STATLER, *Shimoda Story* (1969, reprinted 1986); and GRACE FOX, *Britain and Japan, 1858–1883* (1969). Works on the Meiji Restoration include W.G. BEASLEY, *The Meiji Restoration* (1972); and MARIUS B. JANSEN, *Sakamoto Ryōma and the Meiji Restoration* (1961, reissued 1994). The emergence of the modern state during Meiji is surveyed in WILLIAM W. LOCKWOOD, *The Economic Development of Japan,* expanded ed. (1968, reissued 1974); JOHANNES HIRSCHMEIER, *The Origins of Entrepreneurship in Meiji Japan* (1964); and MARIUS B. JANSEN (ed.), *Changing Japanese Attitudes Toward Modernization* (1965, reissued 1982).

Meiji political issues are the focus of GEORGE AKITA, *Foundations of Constitutional Government in Modern Japan, 1868–1900* (1967); and ROBERT E. WARD (ed.), *Political Development in Modern Japan* (1968, reissued 1973). Foreign relations are treated in HILARY CONROY, *The Japanese Seizure of Korea, 1868–1910* (1960, reissued 1974); JOHN ALBERT WHITE, *The Diplomacy of the Russo-Japanese War* (1964, reissued 1984); and SHUMPEI OKAMOTO, *The Japanese Oligarchy and the Russo-Japanese War* (1970). Meiji social and intellectual trends can be followed in CAROL GLUCK, *Japan's Modern Myths* (1985); and GAIL LEE BERNSTEIN (ed.), *Recreating Japanese Women, 1600–1945* (1991). Developments in agriculture and industry include RICHARD J. SMETHURST, *Agricultural Development and Tenancy Disputes in Japan, 1870–1940* (1986); and SHELDON GARON, *The State and Labor in Modern Japan* (1987).

Works dealing with the emergence of politics and nationalism include TETSUO NAJITA, *Hara Kei in the Politics of Compromise, 1905–1915* (1967); PETER DUUS, *Party Rivalry and Political Change in Taishō Japan* (1968); ROBERT A. SCALAPINO, *Democracy and the Party Movement in Prewar Japan* (1953, reissued 1975); WILLIAM MILES FLETCHER III, *The Search for a New Order* (1982); AKIRA IRIYE, *After Imperialism* (1965, reissued 1990); RAMON H. MYERS and MARK R. PEATTIE (eds.), *The Japanese Colonial Empire, 1895–1945* (1984); JAMES B. CROWLEY, *Japan's Quest for Autonomy* (1966); FRANCIS CLIFFORD JONES, *Japan's New Order in East Asia* (1954, reprinted 1978); and RICHARD STORRY, *The Double Patriots* (1957, reprinted 1973). Studies of the wartime period include RONALD H. SPECTOR, *Eagle Against the Sun* (1985); ROBERT J.C. BUTOW, *Japan's Decision to Surrender* (1954, reissued 1967); THOMAS R.H. HAVENS, *Valley of Darkness: The Japanese People and World War Two* (1978, reprinted 1986); and JOHN W. DOWER, *War Without Mercy* (1986, reissued 1993).

Important postwar studies include MASATAKA KOSAKA, *100 Million Japanese* (1972, reissued as *A History of Postwar Japan,* 1982); ANDREW GORDON (ed.), *Postwar Japan as History* (1993); ROBERT E. WARD and SAKAMOTO YOSHIKAZU (eds.), *Democratizing Japan* (1987); EZRA F. VOGEL, *Japan as Number One* (1979, reprinted 1985); YASUTAKE MURAKAMI and HUGH PATRICK (eds.), *The Political Economy of Japan,* 3 vol. (1987–92); J.A.A. STOCKWIN, *Japan: Divided Politics in a Growth Economy,* 2nd ed. (1982); T.J. PEMPEL, *Policy and Politics in Japan* (1982); I.M. DESTLER *et al., Managing an Alliance: The Politics of U.S.-Japanese Relations* (1976); JOYCE LEBRA, JOY PAULSON, and ELIZABETH POWERS (eds.), *Women in Changing Japan* (1976); MARTIN E. WEINSTEIN, *Japan's Postwar Defense Policy, 1947–1968* (1971); JOHN OWEN HALEY, *Authority Without Power: Law and the Japanese Paradox* (1991); FRANK K. UPHAM, *Law and Social Change in Postwar Japan* (1987); CHAE-JIN LEE, *China and Japan* (1984); DENNIS J. ENCARNATION, *Rivals Beyond Trade: America Versus Japan in Global Competition* (1992); and EDWARD J. LINCOLN, *Japan's New Global Role* (1993). (G.C.H.III/F.G.N.)

Japanese Literature

Both in quantity and quality, Japanese literature ranks as one of the major literatures of the world, comparable in age, richness, and volume to English literature, though its course of development has been quite dissimilar. The surviving works comprise a literary tradition extending from the 7th century AD to the present; during all of this time there was never a "dark age" devoid of literary production. Not only do poetry, the novel, and the drama have long histories in Japan, but some literary genres not so highly esteemed in other countries—including diaries, travel accounts, and books of random thoughts—are also prominent. A considerable body of writing by Japanese in the Chinese classical language, of much greater bulk and importance than comparable Latin writings by Englishmen, testifies to the Japanese literary indebtedness to China. Even the writings entirely in Japanese present an extraordinary variety of styles, which cannot be explained merely in terms of the natural evolution of the language. Some styles were patently influenced by the importance of Chinese vocabulary and syntax; but others developed in response to the internal requirements of the various genres, whether the terseness of haiku (a poem in 17 syllables) or the bombast of the dramatic recitation.

This article is divided into the following sections:

GENERAL CONSIDERATIONS

Problems in reading Japanese literature

The difficulties of reading Japanese literature can hardly be exaggerated; even a specialist in one period is likely to have trouble deciphering a work from another period or genre. Japanese style has always favoured ambiguity, and the particles of speech necessary for easy comprehension of a statement are often omitted as unnecessary or as fussily precise. Sometimes the only clue to the subject or object of a sentence is the level of politeness in which the words are couched; for example, the verb *mesu* (meaning "to eat," "to wear," "to ride in a carriage," etc.) designates merely an action performed by a person of quality. In many cases, ready comprehension of a simple sentence depends on a familiarity with the background of a particular period of history. The verb *miru,* "to see," had overtones of "to have an affair with" or even "to marry" during the Heian period in the 10th and 11th centuries, when men were generally able to see women only after they had become intimate. The long period of Japanese isolation in the 17th and 18th centuries also tended to make the literature provincial, or intelligible only to persons sharing a common background; the phrase "some smoke rose noisily" (*kemuri tachisawagite*), for example, was all readers of the late 17th century needed to realize that an author was referring to the Great Fire of 1682 that ravaged the shogunal capital of Edo (the modern city of Tokyo).

Despite the great difficulties arising from such idiosyncrasies of style, Japanese literature of all periods is exceptionally appealing to modern readers, whether read in the original or in translation. Because it is prevailingly subjective and coloured by an emotional rather than an intellectual or moralistic tone, its themes have a universal quality almost unaffected by time. To read a diary by a court lady of the 10th century is still a moving experience, because she described with such honesty and intensity her deepest feelings that the modern reader forgets the chasm of history and changed social customs separating her world from his own.

The "pure" Japanese language, untainted and unfertilized by Chinese influence, contained remarkably few words of an abstract nature. Just as English borrowed such words as morality, honesty, justice, and the like from the Continent, the Japanese borrowed these terms from China; but if the Japanese language was lacking in the vocabulary appropriate to a Confucian essay, it could express almost infinite shadings of emotional content. A Japanese poet who was dissatisfied with the limitations imposed by his native language or who wished to describe unemotional subjects—whether the quiet outing of aged gentlemen to a riverside or the poet's awareness of his insignificance as compared to the grandeur of the universe—naturally turned to writing poetry in Chinese. From the 16th century on, many words that had been excluded from poetry because of their foreign origins or their humble meanings, following the dictates of the codes of poetic diction established in the 10th century, were adopted by the practitioners of the haiku, originally an iconoclastic, popular verse form. For the most part, however, the Japanese writers, far from feeling dissatisfied with the limitations on expression imposed by their language, were convinced that virtuoso perfection in phrasing and an acute refinement of sentiment were more important to poetry than the voicing of intellectually satisfying concepts.

Effect of the language on literary forms

The Japanese language itself also shaped poetic devices and forms. Because it lacks a stress accent, meaningful rhymes (all words end in one of five simple vowels), or quantity, poetry was distinguished from prose mainly in that it consisted of alternating lines of five and seven syllables; however, if the intensity of emotional expression was low, this distinction alone could not save a poem from dropping into prose. The difficulty of maintaining a high level of poetic intensity may account for the preference for short verse forms that could be polished with perfectionist care. But however moving a tanka (verse in 31 syllables) is, it clearly cannot fulfill some of the functions of longer poetic forms; and there are no Japanese equivalents of *Paradise Lost, The Rape of the Lock,* or *Tintern Abbey.* Instead, the poets devoted their efforts to perfecting each syllable of their compositions, expanding the content of a tanka by suggestion and allusion and prizing shadings of tone and diction more than originality or boldness of expression.

The fluid syntax of the prose affected not only style but content as well. Japanese sentences are sometimes of inordinate length, responding to the subjective turnings and twistings of the author's thought; and the writers considered smooth transitions from one statement to the next, rather than structural unity, the mark of excellent prose. The longer works accordingly betray at times a lack of overall structure of the kind associated in the West with Greek concepts of literary form but consist instead of episodes linked chronologically or by other associations. The difficulty experienced by Japanese writers in organizing their impressions and perceptions into sustained works may explain the development of the diary and travel account, genres in which successive days or the successive stages of a journey provide a structure for otherwise unrelated descriptions. Japanese literature contains some of the world's longest novels and plays; but its genius is most strikingly displayed in the shorter works, whether the tanka, the haiku, the Nō plays, or the poetic diaries.

An acute literary sensibility, fostered especially by the traditions of the court, encouraged the creation of "codes" of poetic practice and of a considerable body of criticism, extending back to the 10th century, that was usually composed by the leading poets or dramatists themselves. These codes exerted an inhibiting effect on new forms of literary composition, but they also helped to preserve a distinctively aristocratic tone.

Relation to Chinese and Korean literature

Japanese literature absorbed much direct influence from China, but the characteristic literary works are strikingly dissimilar. The tradition of feminine writing, especially of such introspective works as diaries, gave a colouring to Japanese prose quite unlike the more objective, masculine Chinese writings. Although the Japanese have been criticized for their imitations of Chinese examples (even by some Japanese), the Japanese novel in fact antedates any Chinese novels by centuries; and the theatre developed quite independently. Because the Chinese and Japanese languages are unrelated, the poetry naturally took different forms, although Chinese poetic examples and literary theories were often in the minds of the Japanese poets. Japanese and Korean are probably related languages, but Korean literary influence was negligible, though Koreans served an important function in transmitting Chinese literary and philosophical works to Japan. Poetry and prose written in the Korean language were unknown to the Japanese until relatively modern times.

From the 8th to the 19th century Chinese literature enjoyed greater prestige among educated Japanese than their own; but a love for the Japanese classics, especially those composed at the court in the 10th and 11th centuries, gradually spread among the entire people and influenced literary expression in every form, even the songs and tales composed by humble people totally removed from the aristocratic world portrayed in classical literature.

HISTORY

Origins. The first writing of literature in Japanese was occasioned by influence from China. The Japanese were still comparatively primitive and without writing when, in the first four centuries AD, knowledge of Chinese civilization gradually reached them. They rapidly assimilated much of this civilization, and the Japanese scribes adopted Chinese characters as a system of writing, although an alphabet (if one had been available to them) would have been infinitely better suited to the Japanese language. The characters, first devised to represent Chinese monosyllables, could be used only with great ingenuity to represent the agglutinative forms of the Japanese language. The ultimate results were chaotic, giving rise to one of the most complicated systems of writing ever invented. The use of Chinese characters enormously influenced modes of expression and led to an association between literary composition and calligraphy lasting many centuries.

Early writings. The earliest Japanese texts were written in Chinese because no system of transcribing the sounds and grammatical forms of Japanese had been invented. The oldest known inscription, on a sword that dates from about AD 440, already showed some modification of normal Chinese usage in order to transcribe Japanese names and expressions. The most accurate way of writing Japanese words was by using Chinese characters not for their meanings but for their phonetic values, giving each character a pronunciation approximating that used by the Chinese themselves. In the oldest extant works, the *Kojiki* (712; "Records of Ancient Matters") and *Nihon shoki*, or *Nihon-gi* (720; "Chronicles of Japan"), more than 120 songs, some perhaps dating back to the 5th century AD, are given in phonetic transcription, doubtless because the Japanese attached great importance to the sounds themselves. In these two works, both officially commissioned "histories" of Japan, many sections are written entirely in Chinese; but parts of the *Kojiki* were composed in a complicated mixture of language that made use of the Chinese characters sometimes for their meaning and sometimes for their sound. (D.Ke.)

The Kojiki and Nihon shoki as collections of myths. Most of the surviving Japanese myths are recorded in these two works. They tell of the origin of the ruling class and were apparently aimed at strengthening its authority. Therefore, they are not pure myths but have much political colouring. They are based on two main traditions: the Yamato Cycle, centred around the sun goddess Amaterasu Ōmikami, and the Izumo Cycle, in which the principal character is Susanoo (or Susanowo) no Mikoto, the brother of Amaterasu.

Myths and genealogies

Genealogies and mythological records were kept in Japan, at least from the 6th century AD and probably long before that. By the time of the emperor Temmu (7th century), it became necessary to know the genealogy of all important families in order to establish the position of each in the eight levels of rank and title modelled after the Chinese court system. For this reason, Temmu ordered the compilation of myths and genealogies that finally resulted in the *Kojiki* and *Nihon shoki*. The compilers of these and other early documents had at their disposal not only oral tradition but also documentary sources. A greater variety of sources was available to the compiler of the *Nihon shoki*. While the *Kojiki* is richer in genealogy and myth, the *Nihon shoki* adds a great deal to scholarly understanding of both the history and the myth of early Japan. Its purpose was to give the newly Sinicized court a history that could be compared with the annals of the Chinese.

Cosmology

The purpose of the cosmologies of the *Kojiki* and *Nihon shoki* is to trace the Imperial genealogy back to the foundation of the world. The myths of the Yamato Cycle figure prominently in these cosmologies. In the beginning, the world was a chaotic mass, an ill-defined egg, full of seeds. Gradually, the finer parts became heaven (Yang), the heavier parts earth (Yin). Deities were produced between the two: first, three single deities, and then a series of divine couples. According to the *Nihon shoki*, one of the first three "pure male" gods appeared in the form of a reed that connected heaven and earth. A central foundation was now laid down for the drifting cosmos, and mud and sand accumulated upon it. A stake was driven in, and an inhabitable place was created. Finally, the god Izanagi (He Who Invites) and the goddess Izanami (She Who Invites) appeared. Ordered by their heavenly superiors, they stood on a floating bridge in heaven and stirred the ocean with a spear. When the spear was pulled up, the brine dripping from the tip formed Onogoro, an island that became solid spontaneously. Izanagi and Izanami then descended to this island, met each other by circling around the celestial pillar, discovered each other's sexuality, and began to procreate. After initial failures, they produced the eight islands that now make up Japan. Izanami finally gave birth to the god of fire and died of burns. Raging with anger, Izanagi attacked his son, from whose blood such deities as the god of thunder were born. Other gods were born of Izanami on her deathbed. They presided over metal, earth, and agriculture. In grief, Izanagi pursued Izanami to Yomi (analogous to Hades) and asked her to come back to the land of the living. The goddess replied that she had already eaten food cooked on a stove in Yomi and could not return. In spite of her warning, Izanagi looked at his wife and discovered that her body was infested with maggots. The angry and humiliated goddess then chased Izanagi from the underworld. When he finally reached the upper world, Izanagi blocked the entrance to the underworld with an enormous stone. The goddess then threatened Izanagi, saying that she would kill a thousand people every day. He replied that he would father one thousand and five hundred children for every thousand she killed. After this, Izanagi pronounced the formula of divorce.

Izanagi then returned to this world and purified himself from the miasma of Yomi no Kuni. From the lustral water falling from his left eye was born the sun goddess Amaterasu Ōmikami, ancestress of the Imperial family. From his right eye was born the moon god Tsukiyomi no Mikoto and from his nose, the trickster god Susanoo. Izanagi gave the sun goddess a jewel from a necklace and told her to govern heaven. He entrusted the dominion of night to the moon god. Susanoo was told to govern the sea. According to the *Kojiki*, Susanoo became dissatisfied with his share and ascended to heaven to see his older sister. Amaterasu, fearing his wild behaviour, met him and suggested that they prove their faithfulness to each other by bringing forth children. They agreed to receive a seed from each other, chew it, and spit it away. If gods rather than goddesses were born, it would be taken as a sign of the good faith of the one toward the other. When Susanoo brought forth gods, his faithfulness was recognized, and he was permitted to live in heaven.

Susanoo trickster myth

Susanoo, becoming conceited over his success, began to play the role of a trickster. He scattered excrement over the dining room of Amaterasu, where she was celebrating the ceremony of the first fruits. His worst offense was to fling into Amaterasu's chamber a piebald horse he had "flayed with a backward flaying" (a ritual offense).

Enraged at the pranks of her brother, the sun goddess hid herself in a celestial cave, and darkness filled the heavens and the earth. The gods were at a loss. Finally, they gathered in front of the cave, built a fire, and made cocks crow. They erected a sacred evergreen tree, and from its branches they hung curved beads, mirrors, and cloth offerings. A goddess named Amenouzume no Mikoto then danced half-nude. Amaterasu, hearing the multitudes of gods laughing and applauding, became curious and opened the door of the cave. Seizing the opportunity, a strong-armed god dragged her out of the cave.

The Izumo Cycle

The myths of the Izumo Cycle then begin to appear in the narration. Having angered the heavenly gods and having been banished from heaven, Susanoo descended to Izumo, where he rescued Princess Marvellous Rice Field (Kushiinada Hime) from an eight-headed serpent. He then married the Princess and became the progenitor of the ruling family of Izumo. The most important member of the family of Susanoo was the god Ōkuninushi no Mikoto, the great earth chief, who assumed control of this region before the descent to earth of the descendants of the sun goddess.

Before long, Amaterasu, the leader of the celestial gods—the gods of Izumo were known as earthly gods—asked Ōkuninushi to turn over the land of Izumo, saying that "the land of the plentiful reed-covered plains and fresh rice ears" was to be governed by the descendants of the heavenly gods. After the submission of Izumo, Amaterasu made her grandson Ninigi no Mikoto (*ninigi* is said to represent rice in its maturity) descend to earth. According to the *Nihon shoki,* Amaterasu handed Ninigi some ears of rice from a sacred rice field and told him to raise rice on earth and to worship the celestial gods. The grandson of the sun goddess then descended to the peak of Takachiho (meaning high thousand ears) in Miyazaki, Kyushu. There he married a daughter of the god of the mountain, named Konohana-sakuya Hime (Princess Blossoms of the Trees).

When Ninigi's wife became pregnant and was about to give birth, all in a single night, he demanded proof that the child was his. She accordingly set fire to her room, then safely produced three sons. One of them, in turn, became the father of the legendary first emperor, Jimmu, who is considered to mark the watershed between the "age of the gods" and the historical age; but Jimmu's eastern expedition and conquest of the Japanese heartland was also a myth. (N.M./D.Ke.)

Origin of the tanka in the Kojiki. The myths in the *Kojiki* are occasionally beguiling, but the only truly literary parts of the work are the songs. The early songs lack a fixed metrical form; the lines, consisting of an indeterminate number of syllables, were strung out to irregular lengths, showing no conception of poetic form. Some songs, however, seem to have been reworked—perhaps when the manuscript was transcribed in the 8th century—into what became the classic Japanese verse form, the tanka (short poem), consisting of five lines of five, seven, five, seven, and seven syllables. Various poetic devices employed in these songs, such as the *makura kotoba* ("pillow word"), a kind of fixed epithet, remained a feature of later poetry.

Altogether, some 500 primitive songs have been preserved in various collections. Many describe travel, and a fascination with place-names, evident in the loving enumeration of mountains, rivers, and towns with their mantic epithets, was developed to great lengths in the gazetteers (*fudoki*) compiled at the beginning of the 8th century. These works, of only intermittent literary interest, devote considerable attention to the folk origins of different place-names, as well as to other local legends.

The significance of the Man'yōshū. A magnificent anthology of poetry, the *Man'yōshū* (compiled after 759; "Collection of Ten Thousand Leaves"), is the single great literary monument of the Nara period (710–784), although it includes poetry written in the preceding century, if not earlier. Most of the 4,500 or so poems are tankas; but the masterpieces of the *Man'yōshū* are the 260 *chōka* ("long poems"), ranging up to 150 lines in length and cast in the form of alternating lines in five and seven syllables followed by a concluding line in seven syllables. The amplitude of the *chōka* permitted the poets to treat themes impossible within the compass of the tanka—whether the death of a wife or child, the glory of the Imperial family, the discovery of a gold mine in a remote province, or the hardships of military service.

Kakinomoto Hitomaro

The greatest of the *Man'yōshū* poets, Kakinomoto Hitomaro, served as a kind of poet laureate in the late 7th and early 8th centuries, accompanying the sovereigns on their excursions and composing odes of lamentation for deceased members of the Imperial family. Modern scholars have suggested that the *chōka* may have originated as exorcisms of the dead, quieting the ghosts of recently deceased persons by reciting their deeds and promising that they will never be forgotten. Some of Hitomaro's masterpieces describe the glories of princes or princesses he may never have met so convincingly as to transcend any difference between "public" expressions of grief and his private feelings. Hitomaro's *chōka* are unique in Japanese poetry thanks to their superb combination of imagery, syntax, and emotional strength; they are works of truly masculine expression. He showed in his tanka, however, that he was also capable of the evocative, feminine qualities typical of later Japanese poetry.

The *chōka* often concluded with one or more *hanka* ("envoys") that resumed central points of the preceding poem. The *hanka* written by the 8th-century poet Yamabe Akahito are so perfectly conceived as to make the *chōka* they follow at times seem unnecessary; the concision and evocativeness of these poems, identical in form with the tanka, are close to the ideals of later Japanese poetry. Nevertheless, the supreme works of the *Man'yōshū* are the *chōka* of Hitomaro, Ōtomo Tabito, Ōtomo Yakamochi (probably the chief compiler of the anthology), and Yamanoue Okura. The most striking quality of the *Man'yōshū* is its powerful sincerity of expression. The poets were certainly not artless songsmiths exclaiming in wonder over the beauties of nature, a picture that is often painted of them by sentimental critics; but their emotions were stronger and more directly expressed than in later poetry. The corpse of an unknown traveller, rather than the falling of the cherry blossoms, stirred in Hitomaro an awareness of the uncertainty of human life.

The *Man'yōshū* is exceptional in the number of poems composed outside the court, whether by frontier guards or persons of humble occupation. Perhaps some of these poems were actually written by courtiers in the guise of commoners, but the use of dialect and familiar imagery contrasts with the strict poetic diction imposed in the 10th century. The diversity of themes and poetic forms also distinguishes the *Man'yōshū* from the more polished but narrower verse of later times. In Okura's famous "Dialogue on Poverty," for example, two men—one poor and the other destitute—describe their miserable lots, revealing a concern over social conditions that would be absent from the classical tanka. Okura's visit to China early in the 8th century, as the member of a Japanese embassy, may account for Chinese influence in his poetry. His poems are also prefaced in many instances by passages in Chinese stating the circumstances of the poems or citing Buddhist parallels.

The *Man'yōshū* was transcribed in an almost perversely complicated system that used Chinese characters arbitrarily, sometimes for meaning and sometimes for sound. The lack of a suitable script probably inhibited literary production in Japanese during the Nara period. The growing importance, however, of Chinese poetry as the mark of literary accomplishment in a courtier may also have interrupted the development of Japanese literature after its first flowering in the *Man'yōshū.*

Eighteen *Man'yōshū* poets are represented in the collection *Kaifūsō* (751), an anthology of poetry in Chinese composed by members of the court. These poems are little more than pastiches of ideas and images borrowed directly from China; the composition of such poetry reflects the enormous prestige of Chinese civilization at this time.

Classical literature: Heian period (794–1185). The foundation of the city of Heian-kyō (later known as Kyōto) as the capital of Japan marked the beginning of a period of great literary brilliance. The earliest writings of the period, however, were almost all in Chinese because of the continued desire to emulate the culture of the continent. Three Imperially sponsored anthologies of Chinese poetry appeared between 814 and 827, and it seemed for a time that writing in Japanese would be relegated to an extremely minor position. The most distinguished writer of Chinese verse, the 9th-century poet Sugawara Michizane, gave a final lustre to this period of Chinese learning by his erudition and poetic gifts; but his refusal to go to China when offered the post of ambassador, on the grounds that China no longer had anything to teach Japan, marked a turning point in the response to Chinese influence.

Poetry. The invention of the *kana* phonetic syllabary, traditionally attributed to the 9th-century Shingon priest and Sanskrit scholar Kūkai, enormously facilitated writing in Japanese. Private collections of poetry in *kana* began to be compiled about 880; and in 905 the *Kokinshū* ("Collection from Ancient and Modern Times"), the first major work of *kana* literature, was compiled by the poet Ki Tsurayuki and others. This anthology contains 1,111 poems divided into 20 books arranged by topics, including six books of seasonal poems, five books of love poems, and single books devoted to such subjects as travel, mourning, and congratulations. The two prefaces are clearly indebted to the theories of poetry described by the compilers of such Chinese anthologies as the *Shih Ching* and *Wen hsüan*, but the preferences they express would be shared by most tanka poets for the next 1,000 years. The preface by Tsurayuki, the oldest work of sustained prose in *kana*, enumerated the circumstances that move men to write poetry; he believed that melancholy, whether aroused by a change in the seasons or by a glimpse of white hairs reflected in a mirror, provided a more congenial mood for writing poetry than the harsher emotions treated in the *Man'yōshū*. The best tanka in the *Kokinshū* captivate the reader by their perceptivity and tonal beauty, but these flawlessly turned miniatures obviously lack the variety of the *Man'yōshū*.

The first major work in *kana*

Skill in composing tanka became an asset in gaining preference at court; it was also essential to a lover, whose messages to his mistress (who presumably could not read Chinese, still the language employed by men in official documents) often consisted of poems describing his own emotions or begging her favours. In this period the tanka almost completely ousted the *chōka* because the shorter poems were more suited to the lover's billet-doux or to competitions on prescribed themes.

For the poets of the *Kokinshū* and the later court anthologies, originality was less desirable than perfection of language and tone. The critics, far from praising novelty of effects, condemned deviations from the standard poetic diction (established by the *Kokinshū*) of some 2,000 words and insisted on absolute adherence to the poetic conventions. Although these restrictions saved Japanese poetry from lapses into bad taste or vulgarity, they froze it for centuries in prescribed modes of expression. Only a skilled critic can distinguish a tanka of the 10th century from one of the 18th century. The *Kokinshū* set the precedent for later court anthologies, and a knowledge of its contents was indispensable to all poets as a guide and source of literary allusions.

Love poetry occupies a prominent place in the *Kokinshū*, but the joys of love are seldom celebrated; instead, the poets wrote in the melancholy vein prescribed in the preface, describing the uncertainties before a meeting with the beloved, the pain of parting, or the sad realization that an affair had ended. The invariable perfection of diction, unmarred by any indecorous cry from the heart, may sometimes make one doubt the poet's sincerity. This is not true of the great *Kokinshū* poets of the 9th century—Ono Komachi, Lady Ise, Ariwara Narihira, and Tsurayuki himself—but even Buddhist priests, who presumably had renounced carnal love, wrote love poetry at the court competitions; and it is hard to detect any difference between such poems and those of sincere lovers.

The preface of the *Kokinshū* lists judgments on the principal poets of the collection. This criticism is unsatisfying to a modern reader because it is so terse and unanalytical; but it nevertheless marks a beginning of Japanese poetic criticism, an art that developed impressively during the course of the Heian period.

The literary diary

Prose. Ki Tsurayuki is celebrated also for his *Tosa nikki* (935; *The Tosa Diary*), the account of his homeward journey to Kyōto from the province of Tosa, where he had served as governor. Tsurayuki wrote this diary in Japanese, though men at the time normally kept their diaries in Chinese (perhaps it was in order to escape reproach for adopting this unmanly style that he pretended a woman in the governor's entourage was the author). Events of the journey are interspersed with the poems composed on various occasions. The work is affecting especially because of the repeated, though muted, references to the death of Tsurayuki's daughter in Tosa.

Tosa nikki is the earliest example of a literary diary. Although Tsurayuki pretended that it was written by a woman, the later Heian diarists who wrote in the Japanese language were, in fact, court ladies; their writings include some of the supreme masterpieces of the literature. *Kagerō nikki* (*The Gossamer Years*) describes the life between 954 and 974 of the second wife of Fujiwara Kaneie, a prominent court official. The first volume, related long after the events, is in the manner of an autobiographical novel; even the author confesses that her remembrances are probably tinged with fiction. The second two volumes approach a true diary, with some entries apparently made on the days indicated. The writer (known only as "the mother of Michitsuna") describes, with many touches of self-pity, her unhappy life with her husband. She evidently assumed that readers would sympathize, and often this is the case, though her self-centred complaints are not endearing. In one passage, in which she gloats over the death of a rival's child, her obsession with her own griefs shows to worst advantage; yet her journal is extraordinarily moving precisely because the author dwells exclusively on universally recognizable emotions and omits the details of court life that must have absorbed the men.

Other diaries of the period include the anecdotal *Murasaki Shikibu nikki* ("The Diary of Murasaki Shikibu"; Eng. trans., *Murasaki Shikibu: Her Diary and Poetic Memoirs*), at once an absorbing literary work and a source of information on the court life the author (Murasaki Shikibu) described more romantically in her masterpiece *Genji monogatari* (*c.* 1010; *The Tale of Genji*), and *Izumi Shikibu nikki* (*The Diary of Izumi Shikibu*), which is less a diary than a short story liberally ornamented with poetry.

These "diaries" are closely related in content and form to the *uta monogatari* ("poem tales") that emerged as a literary genre later in the 10th century. *Ise monogatari* (*c.* 980; *Tales of Ise*) consists of 143 episodes, each containing one or more poems and an explanation in prose of the circumstances of composition. The brevity and often the ambiguity of the tanka gave rise to a need for such explanations, and when these explanations became extended or (as in the case of *Ise monogatari*) were interpreted as biographical information about one poet (Ariwara Narihira), they approached the realm of fiction.

Along with the poem tales, there were works of religious or fanciful inspiration going back to *Nihon ryōiki* (822; *Miraculous Stories from the Japanese Buddhist Tradition*), an account of Buddhist miracles in Japan compiled by the priest Kyōkai. These stories, written in Chinese, were probably used as a source of sermons by the priests with the intent of persuading ordinary Japanese, incapable of reading difficult works of theology, that they must lead virtuous lives if they were not to suffer in hell for present misdeeds. No such didactic intent is noticeable in *Taketori monogatari* (10th century; *Tale of the Bamboo Cutter*), a fairy tale about a princess who comes from the moon to dwell on earth in the house of a humble bamboo cutter; the various tests she imposes on her suitors, fantastic though they are, are described with humour and realism.

Early development of the novel

The first lengthy work of fiction, *Utsubo monogatari* ("The Tale of the Hollow Tree"), was apparently written between 970 and 983, although the last chapter may have been written later. This uneven, ill-digested work is of interest chiefly as an amalgam of elements in the poem tales and fairy tales; it contains 986 tanka, and its episodes range from early realism to pure fantasy.

The contrast between this crude work and the sublime *Genji monogatari* is overwhelming. Perhaps the difference is best explained in terms of the feminine traditions of writing, exemplified especially by the diaries, which enabled Murasaki Shikibu to discover depths in her characters unsuspected by the male author of *Utsubo monogatari.* The *Genji monogatari* is the finest work not only of the Heian period but of all Japanese literature and merits being called the first important novel written anywhere in the world. *Genji monogatari* was called a work of *mono no aware* ("a sensitivity to things") by the great 18th-century literary scholar Motoori Norinaga; the hero, Prince Genji, is not remarkable for his martial prowess or his talents as a statesman but as an incomparable lover, sensitive to each of the many women he wins. The story is related in terms of the successive women Genji loves; each of them evokes a different response from this marvellously complex man. The last third of the novel, describing the world after Genji's death, is much darker in tone; and the principal figures, though still impressive, seem no more than fragmentations of the peerless Genji.

The success of *Genji monogatari* was immediate. The author of the touching *Sarashina nikki* (mid-11th century; "Sarashina Diary"; Eng. trans., *As I Crossed a Bridge of Dreams*) described how as a girl she longed to visit the capital so that she might read the entire work (which had been completed some 10 years earlier). Imitations and derivative works based on *Genji monogatari,* especially on the last third of it, continued to be written for centuries, inhibiting the fiction composed by the court society.

Makura no sōshi (*c.* 1000; *The Pillow Book of Sei Shōnagon*) is another masterpiece of the Heian period that should be mentioned with *Genji monogatari.* Japanese critics have often distinguished the *aware* of *Genji monogatari* and the *okashi* of *Makura no sōshi. Aware* meant sensitivity to the tragic implications of a moment or gesture, *okashi* the comic overtones of perhaps the same moment or gesture. The lover's departure at dawn evoked many wistful passages in *Genji monogatari,* but in *Makura no sōshi* the author noted with unsparing exactness the lover's fumbling, ineffectual leave-taking and his lady's irritation. Murasaki Shikibu's *aware* can be traced through later literature—sensitivity always marked the writings of any author in the aristocratic tradition—but Sei Shōnagon's wit belonged to the Heian court alone.

The Heian court society passed its prime by the middle of the 11th century, but it did not collapse for another 100 years. Long after its political power had been usurped by military men, the court retained its prestige as the fountainhead of culture. But in the 12th century, literary works belonging to a quite different tradition began to appear. *Konjaku monogatari* (early 12th century; "Tales of Now and Then"), a massive collection of religious and folktales drawn not only from the Japanese countryside but also from Indian and Chinese sources, described elements of society that had never been treated in the court novels. These stories, though crudely written, provide glimpses of how the common people spoke and behaved in an age marked by warfare and new religious movements. The collection of folk songs *Ryōjin hishō,* compiled in 1179 by the emperor Go-Shirakawa, suggests the vitality of this burgeoning popular culture even as the aristocratic society was being threatened with destruction.

Medieval literature: Kamakura, Muromachi, and Azuchi-Momoyama periods (1192–1600). *Kamakura period (1192–1333).* The warfare of the 12th century brought to undisputed power military men (samurai) whose new regime was based on martial discipline. Though the samurai expressed respect for the old culture, some of them even studying tanka composition with the Kyōto masters, the capital of the country moved to Kamakura. The lowered position of women under this feudalistic government perhaps explains the noticeable diminution in the importance of writings by court ladies; indeed, there was hardly a woman writer of distinction between the 13th and 19th centuries.

Poetry

The court poets, however, remained prolific: 15 Imperially sponsored anthologies were completed between 1188 and 1439; and most of the tanka followed the stereotypes established in earlier literary periods.

The finest of the later anthologies, the *Shin kokinshū* (*c.* 1205), was compiled by Fujiwara Sadaie, or Teika, among others, and is considered by many as the supreme accomplishment in tanka composition. The title of the anthology—"the new *Kokinshū*"—indicates the confidence of the compilers that the poets represented were worthy successors of those in the 905 collection; they included (besides the great Teika himself) Teika's father, Fujiwara Toshinari (Shunzei), the priest Saigyō, and the former emperor Go-Toba. These poets looked beyond the visible world for symbolic meanings. The brilliant colours of landscapes filled with blossoms or reddening leaves gave way to monochrome paintings; the poet, instead of dwelling on the pleasure or grief of an experience, sought in it some deeper meaning he could sense if not fully express. The tastes of Teika especially dominated Japanese poetic sensibility, thanks not only to his poetry and essays on poetry but to his choices of the works of the past most worthy of preservation.

Prose

Teika is credited also with a novel, *Matsura no miya monogatari* ("Tale of Matsura Shrine"). Though it is unfinished and awkwardly constructed, its dreamlike atmosphere lingers in the mind with the overtones of Teika's poetry; dreams of the past were indeed the refuge of the medieval romancers, who modelled their language on the *Genji monogatari,* though it was now archaic, and borrowed their themes and characters from the Heian masterpieces. Stories about wicked stepmothers are fairly common; perhaps the writers, contrasting their neglect with the fabled lives of the Heian courtiers, identified themselves with the maltreated stepdaughters; and the typical happy ending of such stories—the stepdaughter in *Sumiyoshi monogatari* is married to a powerful statesman and her wicked stepmother humiliated—may have been the dream fulfillment of their own hopes.

Various diaries describe travels between Kyōto and the shogun's capital in Kamakura. Courtiers often made this long journey in order to press claims in lawsuits, and they recorded their impressions along the way in the typical mixture of prose and poetry. *Izayoi nikki* ("Diary of the Waning Moon"; Eng. trans. in *Translations from Early Japanese Literature*) tells of a journey made in 1277 by the nun Abutsu. A later autobiographical work that also contains extensive descriptions of travel is the superb *Towazu-gatari* (*c.* 1307; "Uninvited Remarks"; Eng. trans., *The Confessions of Lady Nijō*) by Lady Nijō, a work (discovered only in 1940) that provides a final moment of glory to the long tradition of introspective writing by women at court.

Although these writings in the aristocratic manner preserved much of the manner of Heian literature, works of different character became even more prominent in the medieval period. There are many collections of Buddhist and popular tales, of which the most enjoyable is the *Uji shūi monogatari* (*A Collection of Tales from Uji*)—a compilation over a period of years of some 197 brief stories. Although the incidents described in these tales are often similar to those found in *Konjaku monogatari,* they are told with considerably greater literary skill.

Gunki monogatari established as a literary genre

An even more distinctive literary genre of the period is the *gunki monogatari,* or war tale. The most famous, *Heike monogatari* (*The Tale of the Heike*), was apparently first written at the court about 1220, probably by a nobleman who drew his materials from the accounts recited by priests of the warfare between the Taira (Heike) and the Minamoto (Genji) families in the preceding century. The celebrated opening lines of the work, a declaration of the impermanence of all things, also states the main subject, the rise and fall of the Taira family. The text, apparently at first in three books, was expanded to 12 in the course of time, as the result of being recited with improvisations by priest-entertainers. This oral transmission may account not only for the unusually large number of textual variants but also for the exceptionally musical and dramatic style of the work. Unlike the Heian novelists, who rarely admitted words of Chinese origin into their works, the reciters of the *Heike monogatari* employed the contrasting sounds of the imported words to produce what has been acclaimed as the great classic of Japanese style. Although the work is curiously uneven, effective scenes being followed by dull passages in which the narrator seems to be stressing the factual accuracy of his materials, it is at least intermittently superb; and it provided many later novelists and dramatists with characters and incidents for their works.

Heike monogatari was by no means the earliest literary work describing warfare; and other writings, mainly historical in content, were graced by literary flourishes uncommon in similar Western works. *Ōkagami* (*c.* 1120?; "The Great Mirror"; Eng. trans., *Ōkagami*), the most famous of the "mirrors" of Japanese history, undoubtedly influenced the composition of *Heike monogatari,* especially in its moralistic tone. *Hōgen monogatari* (Eng. trans., *Hōgen monogatari*) and *Heiji monogatari* (partial Eng. trans. in *Translations from Early Japanese Literature*) chronicle warfare that antedates the events described in *Heike monogatari* but were probably written somewhat later.

War tales continued to be composed throughout the medieval period. The *Taiheiki* ("Chronicle of the Great Peace"; Eng. trans., *Taiheiki*), for example, covers about 50 years, beginning in 1318, when the emperor Go-Daigo ascended the throne. Though revered as a classic by generations of Japanese, it possesses comparatively little appeal for Western readers, no doubt because so few of the figures come alive.

Characters are more vividly described in two historical romances of the mid- to late 14th century, *Soga monogatari,* an account of the vendetta carried out by the Soga brothers, and *Gikeiki* ("Chronicle of Gikei"; Eng. trans., *Yoshitsune*), describing the life of Minamoto Yoshitsune. Though inartistically composed, these portraits of resourceful and daring heroes caught the imaginations of the Japanese; and their exploits are still prominent on the Kabuki stage.

Another important variety of medieval literature was the reflective essays of Buddhist priests. *Hōjō-ki* (1212; *The Ten Foot Square Hut*) by Kamo Chōmei is a hermit's description of his disenchantment with the world and his discovery of peace in a lonely retreat. The elegiac beauty of its language gives this work, brief though it is, the dignity of a classic. Chōmei was also a distinguished poet, and his essay *Mumyōshō* (*c.* 1210–12; "Nameless Notes") is perhaps the finest example of traditional Japanese poetic criticism.

A later priest, Yoshida Kenkō, writing during the days of warfare and unrest that brought an end to the Kamakura shogunate in 1333, the brief restoration of Imperial authority under the emperor Go-Daigo from 1333 to 1335, and the institution of the Ashikaga shogunate in 1338, barely hints at the turmoil of the times in his masterpiece *Tsurezure-gusa* (*c.* 1330; *Essays in Idleness*); instead, he looks back nostalgically to the past, seeking out the survivals of happier days. Kenkō's aesthetic judgments, often based on a this-worldly awareness rather surprising in a Buddhist priest, gained wide currency, especially after the 17th century, when *Tsurezure-gusa* was widely read.

The Muromachi (1338–1573) and Azuchi-Momoyama (1574–1600) periods. In the 15th century a poetic form of plebeian origins displaced the tanka as the preferred medium of the leading poets. Renga (linked verse) had begun as the composition of a single tanka by two people and was a popular pastime even in remote rural areas. One person would compose the first three lines of a tanka, often giving obscure or even contradictory details in order to make it harder for the second person to complete the poem intelligibly. Gradually, renga spread to the court poets, who saw the artistic possibilities of this diversion and drew up "codes" intended to establish renga as an art. These codes made possible the masterpieces of the 15th century, but their insistence on formalities (*e.g.,* how often a "link" on the moon might appear in 100 links, and which links must end with a noun and which with a verb) inevitably diluted the vigour and freshness of the early renga, itself a reaction against the excessively formal tanka. Nevertheless, the renga of the great 15th-century master Iio Sōgi and his associates are unique in their shifting lyrical impulses, moving from link to link like successive moments of a landscape seen from a boat, avoiding any illusion that the whole was conceived in one person's mind.

Poetry

The short stories of the 15th and 16th centuries cannot be said to have high literary value. Many still look back to the world of the Heian court, but others introduce folk materials or else elements of the miraculous in the attempt to interest readers who lacked the education to appreciate the conventional literary manner. Even though many promising stories are ruined by absurdities before their course is run, for a few moments they often give unforgettable glimpses of a society torn by disorder. The stories are anonymous, but the authors seem to have been both courtiers and Buddhist priests.

Prose

Unquestionably the finest literary works of the 15th century are the Nō dramas, especially those by Zeami Motokiyo (see EAST ASIAN ARTS). They were written in magnificent poetry (often compared to "brocade" because of the many allusions to the poetry of the past) and were provided with a structure that is at once extremely economical and free. Many are concerned with the Buddhist sin of attachment: an inability to forget his life in this world prevents a dead man from gaining release but forces him to return again and again as a ghost to relive the violence or passion of his former existence. Only prayer and renunciation can bring about deliverance. Zeami's treatises on the art of Nō display extraordinary perceptivity. His stated aims were dramatic conviction and reality, but these ideals meant ultimates to him and not superficial realism. Some Nō plays, it is true, have little symbolic or supernatural content, but the central elements of a typical program of five Nō plays were found in the highly poetic and elusive masterpieces that suggest a world invisible to the eye but evokable by the actors through the beauty of movements and speech. Unhappiness over a world torn by disorder may have led writers to suggest in their works truths that lie too deep for words. This seems to have been the meaning of *yūgen* ("mystery and depth"), the ideal of the Nō plays. Parallel developments occurred in the tea ceremony, the landscape garden, and monochrome painting, all arts that suggest or symbolize rather than state.

Drama

Literature during the Tokugawa period (1603–1867). The restoration of peace and the unification of Japan were achieved in the early 17th century, and for approximately 250 years the Japanese enjoyed almost uninterrupted peace. During the first half of the Tokugawa period, the cities of Kyōto and Ōsaka dominated cultural activity; but from about 1770 Edo (the modern Tokyo) became paramount. From the mid-1630s to the early 1850s Japan was closed, by government decree, to contact with the outside world. Initially, this isolation encouraged the development of indigenous forms of literature; but, eventually, in the virtual absence of fertilizing influence from abroad, it resulted in provincial writing. The adoption of printing in the early 17th century made a popular literature possible. The Japanese had known the art of printing since at least the 8th century, but they had reserved it exclusively for reproducing Buddhist writings. The Japanese classics existed only in manuscript form. It is possible that the demand for copies of literary works was so small that it could be satisfied with manuscripts, costly though they were; or perhaps aesthetic considerations made the Japanese prefer manuscripts in beautiful calligraphy, sometimes embellished with illustrations. Whatever the case, not until 1591 was a nonreligious work printed. About the same time, Portuguese missionaries in Nagasaki were printing books in the Roman alphabet. In 1593, in the wake of the Japanese invasion of Korea, a printing press with movable type was sent as a present to the emperor Go-Yōzei. Printing soon developed into the hobby or extravagance of the rich, and many examples of Japanese literature began to appear in small editions. Commercial publication began

Role of printing in creating a popular literature in 1609; by the 1620s even works of slight literary value were being printed for a public eager for new books.

Poetry

Early Tokugawa period (1603–c. 1770). Poetry underwent many changes during the early part of the Tokugawa period. At first the court poets jealously maintained their monopoly over the tanka, but gradually other men, many of them *kokugakusha* ("scholars of national learning"), changed the course of tanka composition by attempting to restore to the form the simple strength of *Man'yōshū* poetry. The early 18th-century poet Kamo Mabuchi was the best of the neo-*Man'yōshū* school, but his tanka rarely rise above mere competence in the ancient language.

The chief development in poetry during the Tokugawa shogunate was the emergence of the haiku as an important genre. This exceedingly brief form (17 syllables arranged in lines of five, seven, and five syllables) had originated in the hokku, or opening verse of a renga sequence, which had to contain in its three lines mention of the season, the time of day, the dominant features of the landscape, etc., making it almost an independent poem. The hokku became known as the haiku late in the 19th century, when it was entirely divested of its original function of opening a sequence of verse; but today even the 17th-century hokku are usually called haiku.

As early as the 16th century haikai renga, or comic renga, had been composed by way of diversion after an evening of serious renga composition, reverting to the original social, rather than literary, purpose of making linked verse. As so often happened in Japan, however, a new art, born as a reaction to the stultifying practices of an older art, was "discovered," codified, and made respectable by practitioners of the older art, generally at the cost of its freshness and vitality. Matsunaga Teitoku, a conventional 17th-century poet of tanka and renga who revered the old traditions, became almost in spite of himself the mentor of the new movement in comic verse, largely as the result of pressure from his eager disciples. Teitoku brought dignity to the comic renga and made it a demanding medium, rather than the quip of a moment. His haikai were distinguishable from serious renga not by their comic conception but by the presence of a *haigon*—a word of Chinese or recent origins that was normally not tolerated in classical verse.

Inevitably, a reaction arose against Teitoku's formalism. The poets of the Danrin school, headed by Nishiyama Sōin and Ihara Saikaku, insisted that it was pointless to waste months if not years perfecting a sequence of 100 verses. Their ideal was rapid and impromptu composition; and their verses, generally colloquial in diction, were intended to amuse for a moment rather than to last for all time. Saikaku especially excelled at one-man composition of extended sequences; in 1684 he composed the incredible total of 23,500 verses in a single day and night, too fast for the scribes to do more than tally.

The haiku was perfected into a form capable of conveying poetry of the highest quality by Matsuo Bashō. After passing through an apprenticeship in both Teitoku and Danrin schools, Bashō founded a school of his own, insisting that a haiku must contain both a perception of some eternal truth and an element of contemporaneity, combining the characteristic features of the two earlier schools. Despite their brief compass, Bashō's haiku often suggest, by means of the few essential elements he presents, the whole world from which they have been extracted; the reader must participate in the creation of the poem. Bashō's best known works are travel accounts interspersed with his verses; of these, *Oku no hosomichi* (1694; *The Narrow Road Through the Deep North*) is perhaps the most popular and revered work of Tokugawa literature.

Prose

The general name for the prose composed between 1600 and 1682 is *kana-zōshi*, or "*kana* books," the name originally having been used to distinguish popular writings in the Japanese syllabary from more learned works in Chinese. The genre embraced not only fiction but also works of a near historical nature, pious tracts, books of practical information, guidebooks, evaluations of courtesans and actors, and miscellaneous essays. Only one writer of any distinction is associated with the *kana-zōshi*—Asai Ryōi, a samurai who became the first popular and professional writer in Japanese history. Thanks to the development of relatively cheap methods of printing and a marked increase in the reading public, Ryōi was able to make a living as a writer. Although some of his works are Buddhist, he wrote in a simple style, mainly in *kana*. His most famous novel, *Ukiyo monogatari* (*c.* 1661; "Tales of the Floating World"), is primitive both in technique and in plot; but under his mask of frivolity Ryōi attempted to treat the hardships of a society where the officially proclaimed Confucian philosophy concealed the gross inequalities in the lots of different men.

The first important novelist of the new era was Ihara Saikaku. Some Japanese critics rank him second only to Murasaki Shikibu in all Japanese literature, and his works have been edited with the care accorded only to great classics. Such attention would surely have surprised Saikaku, whose fiction was dashed off almost as rapidly as his legendary performances of comic renga, with little concern for the judgments of posterity.

Saikaku's first novel, *Kōshoku ichidai otoko* (1682; *The Life of an Amorous Man*), changed the course of Japanese fiction. The title itself had strong erotic overtones, and the plot describes the adventures of one man, from his precocious essays at lovemaking as a child of seven to his decision at the age of 60 to sail to an island populated only by women. The licensed quarters of prostitution established in various Japanese cities by the Tokugawa government (despite its professions of Confucian morality), in order to help control unruly samurai by dissipating their energies, became a centre of the new culture. Expertise in the customs of the brothels was judged the mark of the man of the world. The old term *ukiyo*, which had formerly meant the "sad world" of Buddhist stories, now came to designate its homonym, the "floating world" of pleasure; this was the chosen world of Saikaku's hero, Yonosuke, who became the emblematic figure of the era.

Saikaku's masterpiece, *Kōshoku gonin onna* (1686; *Five Women Who Loved Love*), described the loves of women of the merchant class, rather than prostitutes; this was the first time that women of this class were given such attention. In other works he described, sometimes with humour but sometimes with bitterness, the struggles of merchants to make fortunes. His combination of a glittering style and warm sympathy for the characters lifted his tales from the borders of pornography to high art.

Saikaku was a central figure in the renaissance of literature of the late 17th century. The name Genroku (an era name designating the period 1688–1703) is often used of the characteristic artistic products: the Ukiyo-e ("pictures of the floating world"); the *ukiyo-zōshi* ("tales of the floating world"); the Kabuki and *jōruri*, or puppet theatres; and haiku poetry. Unlike its antecedents, this culture prized modernity above conformity to the ancient traditions; to be abreast of the floating world was to be up-to-date, sharing in the latest fashions and slang, delighting in the moment rather than in the eternal truths of Nō plays of medieval poetry.

Another, darker side to Genroku culture is depicted in Saikaku's late works, with their descriptions of the desperate expedients to which men turned in order to pay their bills. Saikaku seldom showed much sympathy for the prostitutes he described; but the chief dramatist of the time, Chikamatsu Monzaemon, wrote his best plays about unhappy women, driven by poverty into their lives as prostitutes, whose only release from the sordid world in which they were condemned to dwell came when they joined their lovers in double suicides. In the world of merchants treated by Chikamatsu, a lack of money, rather than the cosmic griefs of the Nō plays, drove men to death with the prostitutes they loved but could not afford to buy.

Drama

Chikamatsu wrote most of his plays for the puppet theatre, which, in the 18th century, enjoyed even greater popularity than Kabuki. His plays fell into two main categories: those based, however loosely, on historical facts or legends, and those dealing with contemporary life. The domestic plays are rated much higher critically because they avoid the bombast and fantastic displays of heroism that mark the historical dramas; but the latter, adapted for the Kabuki theatre, are superb acting vehicles.

The mainstays of the puppet theatre were written not by Chikamatsu but by his successors; his plays, despite their literary superiority, failed to satisfy the audiences' craving for displays of puppet techniques and for extreme representations of loyalty, self-sacrifice, and other virtues of the society. The most popular puppet play (later adapted also for the Kabuki actors) was *Chūshingura* (1748; "The Treasury of Loyal Retainers"; Eng. trans., *Chūshingura*) by Takeda Izumo and his collaborators; the same men were responsible for half a dozen other perennial favourites of the Japanese stage. The last great 18th-century writer of puppet plays, Chikamatsu Hanji, was a master of highly dramatic, if implausible, plots.

Poetry

Late Tokugawa period (c. 1770–1867). The literature of the late Tokugawa period is generally inferior to earlier achievements, especially those of the Genroku masters. Authentic new voices, however, were heard in traditional poetic forms. Later neo-*Man'yōshū* poets such as Ryōkan, Ōkuma Kotomichi, and Tachibana Akemi proved that the tanka was not limited to descriptions of the sights of nature or disappointed love but could express joy over fish for dinner or wrath at political events. Some poets who felt that tanka did not provide ample scope for the display of such emotions turned, as in the past, to writing poetry in Chinese. The early 19th-century poet Rai Sanyō probably wrote verse in Chinese more skillfully than any previous Japanese.

Later Tokugawa poets also added distinctive notes of their own to the haiku. Yosa Buson, for example, introduced a romantic and narrative element, and Kobayashi Issa employed the accents of the common people.

Prose

A great variety of fiction was produced during the last century of the Tokugawa shogunate, but it is commonly lumped together under the somewhat derogatory heading of *gesaku* ("playful composition"). The word "playful" did not necessarily refer to the subject matter but to the professed attitude of the authors, educated men who disclaimed responsibility for their compositions. Ueda Akinari, the last master of fiction of the 18th century, won a high place in literary history mainly through his brilliant style, displayed to best advantage in *Ugetsu monogatari* (1776; *Tales of Moonlight and Rain*), a collection of supernatural tales. The *gesaku* writers, however, did not follow Akinari in his perfectionist attention to style and construction; instead, they produced books of almost formless gossip, substituting the raciness of daily speech for the elegance of the classical language, and relying heavily on the copious illustrations for success with the public.

The *gesaku* writers were professionals who made their living by sale of their books. They aimed at as wide a public as possible, and when a book was successful it was usually followed by as many sequels as the public would accept. The most popular of the comic variety of *gesaku* fiction was *Tōkai dōchū hizakurige* (1802–22; "Travels on Foot on the Tōkaidō"; Eng. trans., *Shank's Mare*), by Jippensha Ikku, an account of the travels and comic misfortunes of two irrepressible men from Edo along the Tōkaidō, the great highway between Kyōto and Edo. *Shunshoku umegoyomi* (1832–33; "Spring Colours: The Plum Calendar"), by Tamenaga Shunsui, is the story of Tanjirō, a peerlessly handsome but ineffectual young man for whose affections various women fight. The author at one point defended himself against charges of immorality: "Even though the women I portray may seem immoral, they are all imbued with deep sentiments of chastity and fidelity." It was the standard practice of *gesaku* writers, no matter how frivolous their compositions might be, to pretend that their intent was didactic.

The *yomihon* ("books for reading"—so called to distinguish them from works enjoyed mainly for their illustrations) were much more openly moralistic. Although they were considered to be *gesaku,* no less than the most trivial books of gossip, their plots were burdened with historical materials culled from Chinese and Japanese sources, and the authors frequently underlined their didactic purpose. But despite the serious intent of the *yomihon,* they were romances, rather than novels; and their characters, highly schematized, tended to be witches, fairy princesses, and impeccably noble gentlemen. Where they succeeded, as in a few works by Takizawa Bakin, they are absorbing as examples of storytelling rather than as embodiments of the principle of *kanzen chōaku* ("the encouragement of virtue and the chastisement of vice"), Bakin's professed aim in writing fiction.

Japanese literature in general was at one of its lowest levels at the end of the Tokugawa period. A few tanka poets and the Kabuki dramatist Kawatake Mokuami are the only writers of the period whose works are still read today. It was an exhausted literature that could be revived only by the introduction of fresh influences from abroad.

Modern literature. Even after the arrival of Commodore Perry's fleet in 1853 and the gradual opening of the country to the West and its influence, there was little noticeable effect on Japanese literature. The long closure of the country and the general sameness of Tokugawa society for decades at a time seems to have atrophied the imaginations of the *gesaku* writers. Even the presence of curiously garbed foreigners, which should have provoked some sort of reaction from authors searching for new materials, at first produced little effect. The *gesaku* writers were oblivious to the changes in Japanese society, and they continued to grind out minor variants on the same hackneyed themes of the preceding 200 years.

It was only after the removal in 1868 of the capital to Edo (renamed Tokyo) and the declaration of the emperor Meiji that he would seek knowledge from the entire world that the *gesaku* writers realized their days of influence were numbered. They soon fell under attack from their old enemies, the Confucian denouncers of immoral books, and also from advocates of the new Western learning. Although the *gesaku* writers responded with satirical pieces and traditional Japanese fiction deriding the new learning, they were helpless to resist the changes transforming the entire society.

Introduction of Western literature. Translations from European languages of nonliterary works began to appear soon after the Meiji Restoration. The most famous example was the translation (1870) of Samuel Smiles's *Self-Help;* it became a kind of bible for ambitious young Japanese eager to emulate Western examples of success. The first important translation of a European novel was *Ernest Maltravers,* by the British novelist Lord Bulwer-Lytton, which appeared in 1879 under the title *Karyū shunwa* ("A Spring Tale of Blossoms and Willows"). The early translations were inaccurate, and the translators unceremoniously deleted any passages they could not understand readily or which they feared might be unintelligible to Japanese readers. They also felt obliged to reassure readers that, despite the foreign names of the characters, the emotions they felt were exactly the same as those of a Japanese.

Early translations and their influence

It did not take long, however, for the translators to discover that European literature possessed qualities unknown in the Japanese writings of the past. The literary scholar Tsubouchi Shōyō was led by his readings in European fiction and criticism to reject didacticism as a legitimate purpose of fiction; he insisted instead on its artistic values. His critical essay *Shōsetsu shinzui* (1885; "The Essence of the Novel") greatly influenced the writing of subsequent fiction not only because of its emphasis on realism as opposed to didacticism but because Tsubouchi, a member of the samurai class, expressed the conviction that novels, hitherto despised by the intellectuals as mere entertainments for women and children, were worthy of even a scholar's attention.

Ukigumo (1887–89; "Floating Cloud"; Eng. trans., *Japan's First Modern Novel*), by Futabatei Shimei, was the first modern Japanese novel. The author was familiar with Russian literature and contemporary Western literary criticism. Futabatei wrote *Ukigumo* in the colloquial, apparently because his readings in Russian literature had convinced him that only the colloquial could suitably be used when describing the writer's own society. Despite Futabatei's success with this experiment, most Japanese writers continued to employ the literary language until the end of the century. This was due, no doubt, to their reluctance to give up the rich heritage of traditional expression in favour of the unadorned modern tongue.

Western influences on poetry. Translations of Western poetry led to the creation of new Japanese literary forms. The pioneer collection, *Shintaishi-shō* (1882; "Selection of Poems in the New Style"), contained not only translations from English but also five original poems by the translators in the poetic genres of the foreign examples. The translators declared that although European poetry had greater variety than Japanese poetry—some poems are rhymed, others unrhymed, some are extremely long, others abrupt—it was invariably written in the language of ordinary speech. The insistence on modern language and the discovery of the many forms available to the poet were not the only lessons learned from European poetry. The translators also made the Japanese public aware of how much of human experience had never been treated in the tanka or haiku forms.

Innumerable Western critics have sarcastically commented on the Japanese proclivity for imitating foreign literary models and on their alleged indifference to their own traditions. It is true that without Russian examples Futabatei could not have written *Ukigumo,* and without English examples such poets as Shimazaki Tōson could not have created modern Japanese poetry; but far from recklessly abandoning their literary heritage, most writers were at great pains to acquaint themselves with their traditional literature. The outstanding novelists of the 1890s—Ozaki Kōyō, Kōda Rohan, Higuchi Ichiyō, and Izumi Kyōka—all read Saikaku and were noticeably influenced by him. Ichiyō's short novel *Takekurabe* (1895; *Growing Up*) described the children of the Yoshiwara quarter of Edo in a realistic manner quite unlike that of the usual stories about prostitutes and their customers, but she used the language of Saikaku for her narration. Izumi Kyōka, though educated partly at a Western mission school, wrote superbly in the vein of late Tokugawa fiction; something of the distant Japanese literary past pervaded even his writings of the 1930s, the final years of his life.

Influence of Saikaku

In poetry, too, the first products of Western influence were comically inept experiments with rhyme and with such unpromising subjects as the principles of sociology. Shimazaki Tōson's "Akikaze no uta" (1896; "Song of the Autumn Wind"), however, is not merely a skillful echo of Shelley but a true picture of a Japanese landscape; and the irregular lines of his poem tend to fall into the traditional pattern of five and seven syllables.

A decade after the works of such English Romantic poets as Shelley and Wordsworth had influenced Japanese poetry, the translations made by Ueda Bin of the French Parnassian and Symbolist poets made an even more powerful impression. Ueda wrote, "The function of symbols is to help create in the reader an emotional state similar to that in the poet's mind; symbols do not necessarily communicate the same conception to everyone." This view was borrowed from the West, but it accorded perfectly with the qualities of tanka.

Because of the ambiguities of traditional Japanese poetic expression, it was natural for a given poem to produce different effects on different readers; the important thing, as in Symbolist poetry, was to communicate the poet's mood. If the Japanese poets of the early 1900s had been urged to avoid contamination by foreign ideas, they would have declared that this was contrary to the spirit of an enlightened age. But when informed that eminent foreign poets preferred ambiguity to clarity, the Japanese responded with double enthusiasm.

Revitalization of tanka and haiku. Even the traditional forms, tanka and haiku, though moribund in 1868, took on new life, thanks largely to the efforts of Masaoka Shiki, a distinguished late 19th-century poet in both forms but of even greater importance as a critic. Yosano Akiko, Ishikawa Takuboku, and Saitō Mokichi were probably the most successful practitioners of the new tanka. Yosano Akiko's collection *Midaregami* (1901; *Tangled Hair*) stirred female readers especially, not only because of its lyrical beauty but because Akiko herself seemed to be proclaiming a new age of romantic love. Takuboku emerged in the course of his short life (he died in 1912 at the age of 26) as perhaps the most popular tanka poet of all time. His verses are filled with strikingly individual expressions of his intransigent personality. Saitō Mokichi combined an absorption with *Man'yōshū* stylistics and a professional competence in psychiatry. Despite the austere nature of his poetry, he was recognized for many years as the leading tanka poet. In haiku, Takahama Kyoshi built up a following of poets strong enough to withstand the attacks of critics who declared that the form was inadequate to deal with the problems of modern life. Kyoshi himself eventually decided that the function of haiku was the traditional one of an intuitive apprehension of the beauties of nature; but other haiku poets employed the medium to express entirely unconventional themes.

Most tanka and haiku poets continued to use the classical language, probably because its relative concision permitted them to impart greater content to their verses than modern Japanese permits. Poets of the "new style," therefore, were readier to employ the colloquial. Hagiwara Sakutarō, generally considered the finest Japanese poet of the 20th century, brilliantly exploited the musical and expressive possibilities of the modern tongue. Other poets, such as Horiguchi Daigaku, devoted themselves mainly to translations of European poetry, achieving results so compelling in Japanese that these translations are considered to form an important part of the modern poetry of Japan.

"New-style" poetry

The novel between 1905 and 1941. The dominant stream in Japanese fiction since the publication of *Hakai* (1906; *The Broken Commandment*), by Shimazaki Tōson, and *Futon* (1907; *The Quilt*), by Tayama Katai, has been naturalism. Although the movement was originally inspired by the works of the 19th-century French novelist Émile Zola and other European naturalists, it quickly took on a distinctively Japanese colouring, rejecting (as a Confucian scholar might have rejected *gesaku* fiction) carefully developed plots or stylistic beauty in favour of absolute verisimilitude in the author's confessions or in his minute descriptions of the lives of unimportant people hemmed in by circumstances beyond their control.

By general consent, however, the two outstanding novelists of the early 20th century were men who stood outside the naturalist movement, Mori Ōgai and Natsume Sōseki. Ōgai began as a writer of autobiographical fiction with strong overtones of German Romantic writings. Midway in his career he shifted to historical novels that are virtually devoid of fictional elements but are given literary distinction by their concise and masculine style. Sōseki gained fame with humorous novels such as *Botchan* (1906; "The Young Master"; Eng. trans., *Botchan*), a fictionalized account of his experiences as a teacher in a provincial town. *Botchan* has enjoyed phenomenal popularity ever since it first appeared. It is the most approachable of Sōseki's novels, and the Japanese have found pleasure in identifying themselves with the impetuous, reckless, yet basically decent hero. The coloration of Sōseki's subsequent novels became progressively darker, but even the most gloomy have maintained their reputation among Japanese readers, who take it for granted that Sōseki is the greatest of the modern Japanese novelists and who find echoes in their own lives of the mental suffering he described. Sōseki wrote mainly about intellectuals living in a Japan that had been brutally thrust into the 20th century. His best known novel, *Kokoro* (1914; "The Heart"; Eng. trans., *Kokoro*), revolves around another familiar situation in his novels, two men in love with the same woman. His last novel, *Meian* (1916; *Light and Darkness*), though unfinished, has been acclaimed by some as his masterpiece.

An amazing burst of creative activity occurred in the decade following the end of the Russo-Japanese War in 1905. Probably never before in the history of Japanese literature were so many important writers working at once. Three novelists who first emerged into prominence at this time were Nagai Kafū, Tanizaki Jun'ichirō, and Akutagawa Ryūnosuke. Nagai Kafū was infatuated with French culture and described with contempt the meretricious surface of modern Japan. In later years, however, though still alienated from the Japanese present, he showed nostalgia for the Japan of his youth, and his most appealing works contain evocations of the traces of an old and genuine Japan that survived in the parody of Western culture that was Tokyo.

Tanizaki Jun'ichirō

Tanizaki's novels, notably *Tade kuu mushi* (1928–29; "Insects That Eat Knotweed"; Eng. trans., *Some Prefer Nettles*), often presented a conflict between traditional Japanese and Western-inspired ways. In his early works he also proclaimed a preference for the West. Tanizaki's views changed after he moved to the Kansai region in the wake of the Great Kantō Earthquake of 1923, and his subsequent writings traced his gradual accommodation with the old culture of Japan that he had previously rejected. Between 1939 and 1941 Tanizaki published the first of his three modern-language versions of *Genji monogatari.* He willingly sacrificed years of his career to this task because of his unbounded admiration for the supreme work of Japanese literature.

Tanizaki's longest novel, *Sasameyuki* (1943–48; "A Light Snowfall"; Eng. trans., *The Makioka Sisters*), evoked with evident nostalgia the Japan of the 1930s, when people were preoccupied not with the prosecution of a war but with marriage arrangements, visits to sites famous for their cherry blossoms, or the cultural differences between Tokyo and Ōsaka. Two postwar novels by Tanizaki enjoyed great popularity, *Kagi* (1956; *The Key*), the account of a professor's determination to have his fill of sex with his wife before impotence overtakes him; and *Fūten rōjin nikki* (1961–62; *Diary of a Mad Old Man*), a work in a comic vein that describes a very old man's infatuation with his daughter-in-law. No reader would turn to Tanizaki for wisdom as to how to lead his life, nor for a penetrating analysis of society, but his works not only provide the pleasures of well-told stories but also convey the special phenomenon of adulation and rejection of the West that is so prominent a part of the Japanese culture of the 20th century.

Akutagawa established his reputation as a brilliant storyteller who transformed materials found in old Japanese collections by infusing them with modern psychology. No writer enjoyed a greater following in his time, but Akutagawa found less and less satisfaction in his reworkings of existing tales and turned eventually to writing about himself in a sometimes harrowing manner. His suicide in 1927 shocked the entire Japanese literary world. The exact cause is unknown—he wrote of a "vague malaise"—but perhaps Akutagawa felt incapable either of sublimating his personal experiences into pure literature or else of giving them the accents of the proletarian literature movement, then at its height.

The proletarian literature movement in Japan, as in various other countries, attempted to use literature as a weapon to effect reform and even revolution in response to social injustices. Although the movement gained virtual control of the Japanese literary world in the late 1920s, governmental repression beginning in 1928 eventually destroyed it. The chief proletarian writer, Kobayashi Takiji, was tortured to death by the police in 1933. Few of the writings produced by the movement are of literary worth, but the concern for classes of people who had formerly been neglected by Japanese writers gave these works their special significance.

Other writers of the period, convinced that the essential function of literature was artistic and not propagandistic, formed schools such as the "Neo-sensualists" led by Yokomitsu Riichi and Kawabata Yasunari. Yokomitsu's politics eventually moved far to the right, and the promulgation of these views, rather than his efforts to achieve modernism, coloured his later writings; but Kawabata's works (for which he won the Nobel Prize for Literature in 1968) are still admired for their lyricism and intuitive construction. Though Kawabata began as a modernist and experimented with modernist techniques to the end of his career, he is better known for his portraits of women, whether the geisha of *Yukiguni* (1948; *Snow Country*) or the different women whose lives are concerned with the tea ceremony in *Sembazuru* (1952; *Thousand Cranes*).

Japanese critics have divided the fiction of the prewar period into schools, each usually consisting of one leading writer and his disciples. Probably the most influential author was Shiga Naoya. His characteristic literary form was the "I novel" (*watakushi shōsetsu*), a work that treats autobiographical materials with stylistic beauty and great intelligence but is not remarkable for invention. Shiga's commanding presence caused the I novel to be more respected by most critics than outright works of fiction; but the writings of his disciples are sometimes hardly more than pages torn from a diary, of interest only if the reader is already devoted to the author.

The "I novel"

The postwar novel. The aggressive wars waged by the Japanese militarists in the 1930s inhibited literary production. Censorship became increasingly stringent, and writers were expected to promote the war effort. During the Pacific War of 1941–45 little worthwhile literature appeared. Tanizaki began serial publication of *The Makioka Sisters* in 1943, but publication was halted by official order, and the completed work appeared only after the war. The immediate postwar years signalled an extraordinary period of activity, both by the older generation and by new writers. The period is vividly described in the writings of Dazai Osamu, notably *Shayō* (1947; *The Setting Sun*). Other writers described the horrors of the war years; perhaps the most powerful was *Nobi* (1951; *Fires on the Plain*) by Ōoka Shōhei, which described defeated Japanese soldiers in the Philippine jungles. The atomic bombs also inspired much poetry and prose, though it was often too close to the events to achieve artistic integrity. A few works, especially *Kuroi ame* (1965; *Black Rain*) by Ibuse Masuji, succeeded in suggesting the ultimately indescribable horror of the disaster.

Mishima Yukio

The Japan of the immediate postwar period and the prosperous Japan of the 1950s and 1960s provided the background for most of the works of Mishima Yukio, an exceptionally brilliant and versatile novelist and playwright who became the first Japanese writer generally known abroad. Mishima's best known works include *Kinkaku-ji* (1956; *The Temple of the Golden Pavilion*), a psychological study, based on an actual incident, of a young monk who burned a famous architectural masterpiece; and *Hōjō no Umi* (1965–70; *The Sea of Fertility*), the tetralogy he completed on the day of his death. Abe Kōbō was notable among modern writers in that he managed, sometimes by resorting to avant-garde techniques, to transcend the particular condition of being a Japanese and to create myths of suffering humanity in such a work as *Suna no onna* (1962; *The Woman in the Dunes*). The special nature of traditional Japanese culture, which made it infertile ground for Christianity in the 16th century, was treated in several moving novels by Endō Shūsaku, notably *Chimmoku* (1966; *Silence*). The novels of Kita Morio were characterized by an attractive streak of humour that provided a welcome contrast to the prevailingly dark tonality of other contemporary Japanese novels. *Nire-ke no hitobito* (1963–64; *The House of Nire*), though based on the careers of his grandfather and father (the poet Saitō Mokichi) was saved by its humour from becoming no more than an I novel. For almost 20 years Ōe Kenzaburō, a novelist of exceptional power, was treated as the youngest writer of importance, but in the 1970s a new generation at last began to appear with a promise of a renewal of the modern Japanese novel.

The modern drama. The modern Japanese theatre also began with translations and adaptations of Western plays. This new theatre originated at the end of the 19th century, when the public was still too much under the influence of Kabuki to appreciate plays without music or dance. Even in the 20th century, a distinguished dramatist such as Kishida Kunio rarely had the opportunity to see his works performed. The development of modern drama no doubt was hampered by the introduction, at about the same time, of motion pictures, which had a much greater appeal for the public. The most successful playwrights of the 1920s and 1930s, such as Mayama Seika, wrote works that, although the products of modern minds, exploited the special talents of Kabuki actors by treating historical themes and by preserving the traditional stage language. Various distinguished writers were attracted from time to time to the theatre, but they were forced to devote their major efforts to writing fiction, if only because they were so badly remunerated for their plays. It was not until after World War II that modern dramas worthy of an international audience were written and staged.

BIBLIOGRAPHY

General works: SHUICHI KATO, *A History of Japanese Literature*, 3 vol. (1979–83; originally published in Japanese, 2 vol., 1975–80), considers Japanese literature as a key to Japanese intellectual history. JIN'ICHI KONISHI, *A History of Japanese Literature*, trans. from the Japanese, ed. by EARL MINER (1984–), gives special attention to relations among the literatures of other countries of Asia. DONALD KEENE (ed.), *Anthology of Japanese Literature to the Nineteenth Century*, rev. ed. (1978), offers selections from earliest times, with introductions. A good but brief discussion of Japanese myths is E. DALE SAUNDERS, "Japanese Mythology," in SAMUEL N. KRAMER (ed.), *Mythologies of the Ancient World* (1961). GEOFFREY BOWNAS and ANTHONY THWAITE (trans.), *The Penguin Book of Japanese Verse* (1964), gives examples in every form. ROBERT H. BROWER and EARL MINER, *Japanese Court Poetry* (1961), is an excellent study extending from the earliest period to the 14th century. BURTON WATSON (trans.), *Japanese Literature in Chinese*, 2 vol. (1975–76), provides an excellent selection of poetry and some examples in prose, and his work with HIROAKI SATO (eds. and trans.), *From the Country of Eight Islands: An Anthology of Japanese Poetry* (1981), contains poetry of every period. EARL MINER (ed.), *Japanese Poetic Diaries* (1969, reprinted 1976), contains examples from the Heian period to the 19th century.

Early and Nara periods: *Kojiki* (1968), and *This Wine of Peace, This Wine of Laughter: A Complete Anthology of Japan's Earliest Songs* (1968), are both translations by DONALD L. PHILIPPI. *The Manyōshū: One Thousand Poems Selected and Translated from the Japanese* (1940, reissued 1965), contains an excellent selection from this great anthology; and IAN HIDEO LEVY, *The Ten Thousand Leaves* (1981–), is the first volume of a complete translation.

Heian period: IVAN MORRIS, *The World of the Shining Prince* (1964, reissued 1979), provides the social and historical background for the Heian literary masterpieces. KYOKO MOTOMOCHI NAKAMURA (trans.), *Miraculous Stories from the Japanese Buddhist Tradition: The Nihon Ryōiki of the Monk Kyōkai*(1973), also contains Buddhisttales from Indian and Chinese sources. *Kokinshū* is a complete translation by LAUREL RASPLICA RODD and MARY CATHERINE HENKENIUS (1984). *The Gossamer Years*, trans. by EDWARD SEIDENSTICKER (1964, reprinted 1973); MURASAKI SHIKIBU, *The Tale of Genji*, trans. by EDWARD SEIDENSTICKER (1976, reprinted 1981), and another translation by ARTHUR WALEY (1926–33); and *The Pillow Book of Sei Shōnagon*, 2 vol. (1967), and *As I Crossed a Bridge of Dreams* (1971), both trans. by IVAN MORRIS, are all poetic renderings of these classics. *The Izumi Shikibu Diary*, trans. by EDWIN A. CRANSTON (1969); *Murasaki Shikibu, Her Diary and Poetic Memoirs*, trans. by RICHARD BOWRING (1982); *Tales of Ise*, trans. by HELEN C. MCCULLOUGH (1968); and *Tales of Yamato*, trans. by MILDRED M. TAHARA (1980), are more literal versions with scholarly introductions. HELEN C. MCCULLOUGH (trans.), *Ōkagami, the Great Mirror* (1980), and, with WILLIAM H. MCCULLOUGH (trans.), *A Tale of Flowering Fortunes: Annals of Japanese Aristocratic Life in the Heian Period* (1980), are histories written with an admixture of poetry and fiction. THOMAS H. ROHLICH (trans.), *A Tale of Eleventh Century Japan: Hamamatsu Chūnagon Monogatari*, is an example of later Heian fiction. MARIAN URY (trans.), *Tales of Times Now Past* (1979), includes 62 stories from the *Konjaku monogatari*. WILLIAM R. LAFLEUR (trans.), *Mirror for the Moon* (1978), is a collection of free but poetic translations of waka by Saigyō.

Middle Ages: WILLIAM R. LAFLEUR, *The Karma of Words: Buddhism and the Literary Arts in Medieval Japan* (1983), provides a general background for the literature of the period. *The Tale of the Heike*, trans. by HIROSHI KITAGAWA and BRUCE T. TSUCHIDA (1975), is a rendering of the classic account of the warfare between the Taira and Minamoto clans. *The Taiheiki* (1959, reprinted 1976), and *Yoshitsune* (1966), both trans. by HELEN C. MCCULLOUGH, are accurate versions of war tales in rather old-fashioned language. DONALD KEENE (trans. and ed.), *Essays in Idleness* (1967), and *Twenty Plays of the Nō Theatre* (1970), are readable though fairly literal; ARTHUR WALEY (trans.), *The Nō Plays of Japan* (1922, reissued 1979), gives freer versions of the plays. D.E. MILLS (trans.), *A Collection of Tales from Uji* (1970), is a readable and scholarly study. KAREN BRAZELL (trans.), *The Confessions of Lady Nijō* (1973), is a diary of exceptional interest. EARL MINER, *Japanese Linked Poetry* (1979), is a study of renga and haikai poetry with translations.

Tokugawa period: DONALD KEENE, *World Within Walls* (1976, reprinted 1978), is a history of the literature of the period. The several excellent translations of works by Ihara Saikaku include *Five Women Who Loved Love*, trans. by W. THEODORE DE BARY (1956, reprinted 1973); *The Japanese Family Storehouse*, trans. by G.W. SARGENT (1959, reprinted 1969); *The Life of an Amorous Woman, and Other Writings*, trans. by IVAN MORRIS (1963, reissued 1969); and *Worldly Mental Calculations*, trans. by BEN BEFU (1976). HOWARD HIBBETT, *The Floating World in Japanese Fiction* (1959, reissued 1975), is a critical study with translations. MAKOTO UEDA, *Matsuo Bashō* (1970, reissued 1982), contains biographical and critical material on the great haiku poet. NOBUYUKI YUASA (trans.), *The Narrow Road to the Deep North, and Other Travel Sketches* (1966), contains all of Bashō's travel accounts. EARL MINER and HIROKO ODAGIRI (trans.), *The Monkey's Straw Raincoat and Other Poetry of the Bashō School* (1981), gives examples of the linked verse. CHIKAMATSU MONZAEMON, *Major Plays* (1961); and TAKEDA IZUMO, MIYOSHI SHŌRAKU, and NAMIKI SENRYŪ, *Chūshingura: The Treasury of Loyal Retainers* (1971), both trans. by DONALD KEENE; TAKEDA IZUMO *et al.*, *Sugawara and the Secrets of Calligraphy*, trans. by STANLEIGH H. JONES, JR. (1985); and MOKUAMI KAWATAKE, *Love of Izayoi and Seishin*, trans. by FRANK T. MOTOFUJI (1966), are representative plays of the Tokugawa period. UEDA AKINARI, *Ugetsu Monogatari: Tales of Moonlight and Rain*, trans. by LEON ZOLBROD (1974), is a collection of stories of the supernatural. *Ryōkan* (1977), and *Grass Hill* (1983), both trans. by BURTON WATSON, contain poems by monks.

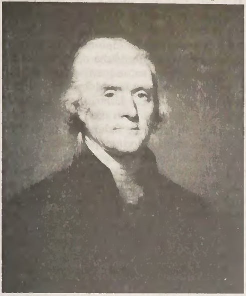
Modern period: DONALD KEENE, *Dawn to the West: Japanese Literature of the Modern Era*, 2 vol. (1984), is a history since 1868. MAKOTO UEDA, *Modern Japanese Writers and the Nature of Literature* (1976), and *Modern Japanese Poets and the Nature of Literature* (1983), are valuable studies. MASAO MIYOSHI, *Accomplices of Silence: The Modern Japanese Novel* (1974, reprinted 1982), is an absorbing study. J. THOMAS RIMER, *Modern Japanese Fiction and Its Tradition: An Introduction* (1978), traces the native elements in modern literature. DONALD KEENE (ed.), *Modern Japanese Literature* (1956); IVAN MORRIS (ed.), *Modern Japanese Stories* (1961, reissued 1970); YUKIO MISHIMA and GEOFFREY BOWNAS (eds.), *New Writing in Japan* (1972); and HOWARD HIBBETT (ed.), *Contemporary Japanese Literature* (1977), are anthologies of modern writing in different genres. TAKAMICHI NINOMIYA and D.J. ENRIGHT (eds.), *The Poetry of Living Japan* (1957, reprinted 1979); ICHIRA KŌNO and RIKUTARO FUKUDA (eds. and trans.), *An Anthology of Modern Japanese Poetry* (1957, reprinted 1971); KENNETH REXROTH and IKUKO ATSUMI (eds. and trans.), *The Burning Heart: Women Poets of Japan* (1977); and *Modern Japanese Poetry*, trans. by JAMES KIRKUP, ed. by A.R. DAVIS (1978), are representative collections. JANINE BEICHMAN, *Masoka Shiki* (1982); and MASAOKA SHIKI, *Peonies Kana: Haiku by the Upasaka Shiki*, trans. and ed. by HAROLD J. ISAACSON (1972), are devoted to the Meiji poet. AMY VLADECK HEINRICH, *Fragments of Rainbows* (1983), is a study of life and poetry of Saitō Mokichi. RICHARD BOWRING, *Mori Ōgai and the Modernization of Japanese Culture* (1979); and J. THOMAS RIMER, *Mori Ōgai* (1975), are studies of this important writer. EDWIN MCCLELLAN, *Two Japanese Novelists* (1969, reissued 1971), discusses major works by Natsume Sōseki and Shimazaki Tōson. EDWARD SEIDENSTICKER, *Kafū the Scribbler* (1965), is a biography of Nagai Kafū with translations from his writings. DENNIS KEENE, *Yokomitsu Riichi, Modernist* (1980), is a study of the chief figure in the Shinkankaku movement. *Modern Japanese Drama*, ed. and trans. by TED T. TAKAYA (1979), contains plays by five outstanding modern dramatists.

(D.Ke.)

Jefferson

Third president of the United States, Thomas Jefferson was born on April 2 [April 13, New Style], 1743, at Shadwell, Virginia. Jefferson was draftsman of the Declaration of Independence, the first U.S. secretary of state (1789–94), and second vice president (1797–1801), and, as president (1801–09), he was responsible for the Louisiana Purchase. An early advocate of total separation of church and state, he also was the founder and architect of the University of Virginia and the most eloquent American proponent of individual freedom as the core meaning of the American Revolution.

By courtesy of the White House Collection, Washington, D.C.



Jefferson, oil painting by Rembrandt Peale, 1800. In the White House Collection, Washington, D.C.

Early years. Albemarle county, where Jefferson was born, lay in the foothills of the Blue Ridge Mountains in what was then regarded as a western province of the Old Dominion. His father, Peter Jefferson, was a self-educated surveyor who amassed a tidy estate that included 60 slaves. According to family lore, Jefferson's earliest memory was as a three-year-old boy "being carried on a pillow by a mounted slave" when the family moved from Shadwell to Tuckahoe. His mother, Jane Randolph Jefferson, was descended from one of the most prominent families in Virginia. She raised two sons, of whom Jefferson was the eldest, and six daughters. There is reason to believe that Jefferson's relationship with his mother was strained, especially after his father died in 1757, because he did everything he could to escape her supervision and had almost nothing to say about her in his memoirs. He boarded with the local schoolmaster to learn his Latin and Greek until 1760, when he entered the College of William and Mary in Williamsburg. By all accounts he was an obsessive student, often spending 15 hours with his books, 3 hours practicing his violin, and the remaining 6 hours eating and sleeping. He read law from 1762 to 1767, then left Williamsburg to practice, mostly representing small-scale planters from the western counties in cases involving land claims and titles.

In 1768 he made two important decisions: first, to build his own home atop an 867-foot- (264-metre-) high mountain near Shadwell that he eventually named Monticello; and second, to stand as a candidate for the House of Burgesses. These decisions nicely embodied the two competing impulses that would persist throughout his life—namely, to combine an active career in politics with periodic seclusion in his own private haven. His political timing was also impeccable, for he entered the Virginia legislature just as opposition to the taxation policies of the British Parliament was congealing.

In the early 1770s his own character was also congealing. In 1772 he married Martha Wayles Skelton, an attractive and delicate young widow whose dowry more than doubled his holdings in land and slaves. In 1774 he wrote "A Summary View of the Rights of British America," which was quickly published, though without his permission, and catapulted him into visibility beyond Virginia as an early advocate of American independence from Parliament's authority; the American colonies were tied to Great Britain, he believed, only by wholly voluntary bonds of loyalty to the king.

His reputation thus enhanced, the Virginia legislature appointed him a delegate to the Second Continental Congress in the spring of 1775. He rode into Philadelphia—and into American history—on June 20, 1775, a tall (slightly above 6 feet 2 inches [1.88 metres]) and gangly young man with reddish blond hair, hazel eyes, a burnished complexion, and rock-ribbed certainty about the American cause. In retrospect, the central paradox of his life was also on display, for the man who the following year was to craft the most famous manifesto for human equality in world history arrived in an ornate carriage drawn by four handsome horses and accompanied by three slaves.

Declaring independence. Jefferson's inveterate shyness prevented him from playing a significant role in the debates within the Congress. John Adams, a leader in those debates, remembered that Jefferson was silent even in committee meetings, though consistently staunch in his support for independence. His chief role was as a draftsman of resolutions. In that capacity, on June 11, 1776, he was appointed to a five-person committee, which also included Adams and Benjamin Franklin, to draft a formal statement of the reasons why a break with Great Britain was justified. Adams asked him to prepare the first draft, which he did within a few days. He later claimed that he was not striving for "originality of principle or sentiment," only seeking to provide "an expression of the American mind"; that is, putting into words those ideas already accepted by a majority of Americans. This accurately describes the longest section of the Declaration of Independence, which lists the grievances against George III. It does not, however, describe the following 55 words, which are generally regarded as the seminal statement of American political culture:

Drafting the Declaration of Independence

> We hold these truths to be self-evident; that all men are created equal; that they are endowed by their Creator with certain inalienable rights; that among these are life, liberty and the pursuit of happiness; that to secure these rights, governments are instituted among men, deriving their just powers from the consent of the governed.

On July 3–4 the Congress debated and edited Jefferson's draft, deleting and revising fully one-fifth of the text. But they made no changes whatsoever in this passage, which over succeeding generations became the lyrical sanction for every liberal movement in American history. At the time, Jefferson himself was disconsolate that the Congress had seen fit to make any changes in his language. Nevertheless, he was not regarded by his contemporaries as the author of the Declaration, which was seen as a collective effort by the entire Congress. Indeed, he was not known by most Americans as the principal author until the 1790s.

He returned to Virginia in October 1776 and immediately launched an extensive project for the reform of the state's legal code to bring it in line with the principles of the American Revolution. Three areas of reform suggest the arc of his political vision: first, he sought and secured abolition of primogeniture, entail, and all those remnants of feudalism that discouraged a broad distribution of property; second, he proposed a comprehensive plan of educational reform designed to assure access at the lowest level for all citizens and state support at the higher levels for the most talented; and third, he advocated a law prohibiting any religious establishment and requiring complete separation of church and state. The last two proposals were bitterly contested, especially the statute for religious freedom, which was not enacted until 1786.

At the end of what was probably the most creative phase of his public career, personal misfortune struck in two successive episodes. Elected governor of Virginia in 1779, he was caught off-guard by a surprise British invasion in 1780 against which the state was defenseless. His flight from approaching British troops was described in the local press, somewhat unfairly, as a cowardly act of abdication. (Critics would recall this awkward moment throughout the remainder of his long career.) Then, in September 1782, his wife died after a difficult delivery in May of their third daughter. These two disasters caused him to vow that he would never again desert his family for his country.

American in Paris. The vow was sincere but short-lived. Jefferson agreed, albeit reluctantly, to serve as a delegate to the Continental Congress in December 1782, where his major contribution was to set forth the principle that territories in the West should not be treated as colonies but rather should enter the Union with status equal to the original states once certain conditions were met. Then, in 1784, recognizing the need to escape the memories of Martha that haunted the hallways at Monticello, he agreed to replace Franklin as American minister to France; or, as legend tells the story, he agreed to succeed Franklin, noting that no one could replace him.

During his five-year sojourn in Paris, Jefferson accomplished very little in any official sense. Several intractable conditions rendered his best diplomatic efforts futile: the United States was heavily in debt owing to the recent war, so few European nations were interested in signing treaties of amity and commerce with the infant American republic; the federal government created under the Articles of Confederation was notoriously weak, so clear foreign policy directives proved impossible; Great Britain already enjoyed a monopoly, controlling more than 80 percent of America's foreign trade, so it had no incentive to negotiate commercial treaties on less favourable terms; and France was drifting toward a cataclysmic political crisis of its own, so relations with the upstart new nation across the Atlantic were hardly a high priority.

But the Paris years were important to Jefferson for personal reasons and are important to biographers and historians for the new light they shed on his famously elusive personality. The dominant pattern would seem to be the capacity to live comfortably with contradiction. For example, he immersed himself wholeheartedly in the art, architecture, wine, and food of Parisian society but warned all prospective American tourists to remain in America so as to avoid the avarice, luxury, and sheer sinfulness of European fleshpots. He made a point of bringing along his elder daughter, Martha (called Patsy as a girl), and later sent for his younger daughter, Maria (called Polly), all as part of his genuine devotion as a single parent. But he then placed both daughters in a convent, wrote them stern lecturelike letters about proper female etiquette, and enforced a patriarchal distance.

Slavery and racism. Even before his departure from France, Jefferson had overseen the publication of *Notes on the State of Virginia*. This book, the only one Jefferson ever published, was part travel guide, part scientific treatise, and part philosophical meditation. Jefferson had written it in 1781 and had agreed to a French edition only after learning that an unauthorized version was already in press.

Discussion in *Notes*

Notes contained an extensive discussion of slavery, including a graphic description of its horrific effects on both blacks and whites, a strong assertion that it violated the principles on which the American Revolution was based, and an apocalyptic prediction that failure to end slavery would lead to "convulsions which will probably never end but in the extermination of one or the other race." It also contained the most explicit assessment that Jefferson ever wrote of what he believed were the biological differences between blacks and whites, an assessment that exposed the deep-rooted racism that he, like most Americans and almost all Virginians of his day, harboured throughout his life.

Sally Hemings

To his critics in later generations, Jefferson's views on race seemed particularly virulent because of his relationship with Sally Hemings, an attractive house slave who bore several children obviously fathered by a white man and some of whom had features resembling those of Jefferson. The public assertion of this relationship was originally made in 1802 by a disreputable journalist interested in injuring Jefferson's political career. His claim was corroborated, however, by one of Sally Hemings's children in an 1873 newspaper interview and then again in a 1968 book by Winthrop Jordan revealing that Hemings became pregnant only when Jefferson was present at Monticello. Finally, in 1998, DNA samples were gathered from living descendants of Jefferson and Hemings. Tests revealed that Jefferson was almost certainly the father of some of Hemings's children. What remained unclear was the character of the relationship—consensual or coercive, a matter of love or rape, or a mutually satisfactory arrangement. Jefferson's admirers preferred to consider it a love affair and to see Jefferson and Hemings as America's preeminent biracial couple. His critics, on the other hand, considered Jefferson a sexual predator whose eloquent statements about human freedom and equality were hypocritical.

In any case, coming as it did at the midpoint of Jefferson's career, the publication of *Notes* affords the opportunity to review Jefferson's previous and subsequent positions on the most volatile and therefore most forbidden topic in the revolutionary era. Early in his career Jefferson had taken a leadership role in pushing slavery onto the political agenda in the Virginia assembly and the federal Congress. In the 1760s and '70s, like most Virginia planters, he endorsed the end of the slave trade. (Virginia's plantations were already well stocked with slaves, so ending the slave trade posed no economic threat and even enhanced the value of the existent slave population.) In his original draft of the Declaration of Independence, he included a passage, subsequently deleted by the Continental Congress, blaming both the slave trade and slavery itself on George III. Unlike most of his fellow Virginians, Jefferson was prepared to acknowledge that slavery was an anomaly in the American republic established in 1776. His two most practical proposals came in the early 1780s: a gradual emancipation scheme by which all slaves born after 1800 would be freed and their owners compensated, and a prohibition of slavery in all the territories of the West as a condition for admission to the Union. By the time of the publication of *Notes*, then, Jefferson's record on slavery placed him among the most progressive elements of southern society. Rather than ask how he could possibly tolerate the persistence of slavery, it is more historically correct to wonder how this member of Virginia's planter class had managed to develop such liberal convictions.

Dating the onset of a long silence is an inevitably imprecise business, but by the time of his return to the United States in 1789 Jefferson had backed away from a leadership position on slavery. The very logic of his argument in *Notes* exposed the inherent intractability of his position. Although he believed that slavery was a gross violation of the principles celebrated in the Declaration of Independence, he also believed that people of African descent were biologically inferior to whites and could never live alongside whites in peace and harmony. They would have to be transported elsewhere, back to Africa or perhaps the Caribbean, after emancipation. Because such a massive deportation was a logistical and economic impossibility, the unavoidable conclusion was that, though slavery was wrong, ending it, at least at present, was inconceivable. That became Jefferson's public position throughout the remainder of his life.

Party politics. Jefferson returned to the United States in 1789 to serve as the first secretary of state under President George Washington. He was entering the most uncharted waters in American history. There had never been an enduring republican government in a nation as large as the United States, and no one was sure if it was possible or how it would work. The Constitution ratified in 1788 was still a work-in-progress, less a blueprint that provided answers than a framework for arguing about the salient questions. And because Jefferson had been serving in France when the constitutional battles of 1787–88 were waged in Philadelphia and then in the state ratifying conventions, he entered the volatile debates of the 1790s without a clear track record of his constitutional convictions.

Role as secretary of state

During his tenure as secretary of state (1790–93), foreign policy was his chief responsibility. Within the cabinet a three-pronged division soon emerged over American policy toward the European powers. While all parties embraced some version of the neutrality doctrine, the specific choices posed by the ongoing competition for supremacy in Europe between England and France produced a bitter conflict. Washington and Adams, who was serving as vice president, insisted on complete neutrality, which in practice meant tacking back and forth between the two dominant world powers of the moment. Alexander Hamilton pushed for a pro-English version of neutrality—chiefly commercial ties with the most potent mercantile power in the world. Jefferson favoured a pro-French version of neutrality, arguing that the Franco-American treaty of 1778 obliged the United States to honour past French support during the war for independence, and that the French Revolution embodied the "spirit of '76" on European soil. Even when the French Revolution spun out of control and began to devour its own partisans, Jefferson insisted that these bloody convulsions were only temporary excesses justified by the larger ideological issues at stake.

This remained his unwavering position throughout the decade. Even after he retired from office late in 1793, he issued directives from Monticello opposing the Neutrality Act (1793) and the Jay Treaty (1795) as pacts with the British harlot and betrayals of our French brethren. Serving as vice president during the Adams presidency (1796–1800), Jefferson worked behind the scenes to undermine Adams's efforts to sustain strict neutrality and blamed the outbreak of the "quasi-war" with France in 1797–98 on what he called "our American Anglophiles" rather than the French Directory. His foreign-policy vision was resolutely moralistic and highly ideological, dominated by a dichotomous view of England as a corrupt and degenerate engine of despotism and France as the enlightened wave of the future.

Jefferson's position on domestic policy during the 1790s was a variation on the same ideological dichotomy. As Hamilton began to construct his extensive financial program—to include funding the national debt, assuming the state debts, and creating a national bank—Jefferson came to regard the consolidation of power at the federal level as a diabolical plot to subvert the true meaning of the American Revolution. As Jefferson saw it, the entire Federalist commitment to an energetic central government with broad powers over the domestic economy replicated the arbitrary policies of Parliament and George III, which the American Revolution had supposedly repudiated as monarchical and aristocratic practices, incompatible with the principles of republicanism. Jefferson sincerely believed that the "principles of '76" were being betrayed by a Federalist version of the "court party," whose covert scheme was to install monarchy and a pseudo-aristocracy of bankers and "monocrats" to rule over the American yeomanry.

Partnership with Madison

The partnership between Jefferson and James Madison, labeled by subsequent historians as "the great collaboration," deserves special attention. John Quincy Adams put it nicely when he observed that "the mutual influence of these two mighty minds on each other is a phenomenon, like the invisible and mysterious movements of the magnet in the physical world." Because the notion of a legitimate opposition to the elected government did not yet exist, and because the term *party* remained an epithet that was synonymous with faction, meaning an organized effort to undermine the public interest, Jefferson and Madison were labeled as traitors by the Federalist press. They were, in effect, inventing a modern form of political behaviour before there was any neutral vocabulary for talking about it. Jefferson's own capacity to live comfortably with contradictions served him well in this context, since he was creating and leading a political party while insisting that parties were evil agents. In 1796 he ran for the presidency against Adams, all the while claiming not to know that he was even a candidate. Most negative assessments of Jefferson's character date from this period, especially the charge of hypocrisy and duplicity.

The highly combustible political culture of the early republic reached a crescendo in the election of 1800, one of the most fiercely contested campaigns in American history. The Federalist press described Jefferson as a pagan and atheist, a treasonable conspirator against the duly elected administrations of Washington and Adams, a utopian dreamer with anarchistic tendencies toward the role of government, and a cunning behind-the-scenes manipulator of Republican propaganda. All these charges were gross exaggerations, save the last. Always operating through intermediaries, Jefferson paid several journalists to libel Adams, his old friend but current political enemy, and offered the vice presidency to Aaron Burr in return for delivering the electoral votes of New York. In the final tally the 12 New York votes made the difference, with the tandem of Jefferson and Burr winning 73 to 65. A quirk in the Constitution, subsequently corrected in the 12th Amendment, prevented electors from distinguishing between their choice of president and vice president, so Jefferson and Burr tied for the top spot, even though voter preference for Jefferson was incontestable. The decision was thrown into the House of Representatives where, after several weeks of debate and backroom wheeling and dealing, Jefferson was elected on the 36th ballot.

Presidency. As Jefferson rose to deliver his inaugural address on March 4, 1801, in the still-unfinished Capitol of the equally unfinished national capital on the Potomac, the mood was apprehensive. The most rabid alarmists had already been proved wrong, since the first transfer of power from one political regime to another had occurred peacefully, even routinely. But it was still very much an open question whether, as Lincoln later put it, "any nation so conceived and so dedicated could long endure" in the absence of a central government along Federalist lines.

Inaugural address

The major message of Jefferson's inaugural address was conciliatory. Its most famous line ("We are all republicans—we are all federalists") suggested that the scatological party battles of the previous decade must cease. He described his election as a recovery of the original intentions of the American Revolution, this after the hostile takeover of those "ancient and sacred truths" by the Federalists, who had erroneously assumed that a stable American nation required a powerful central government. In Jefferson's truly distinctive and original formulation, the coherence of the American republic did not require the mechanisms of a powerful state to survive or flourish. Indeed, the health of the emerging American nation was inversely proportional to the power of the federal government, for in the end the sovereign source of republican government was voluntary popular opinion, "the people," and the latent energies these liberated individuals released when unburdened by government restrictions.

Initially, at least, his policies as president reflected these priorities, which meant dismantling the embryonic federal government, the army and navy, and all federal taxation programs, as well as placing the national debt, which stood at $112 million, on the road to extinction. These reforms enjoyed considerable success for two reasons. First, the temporary cessation of the war between England and France for European supremacy permitted American merchants to trade with both sides and produced unprecedented national prosperity. Second, in selecting Albert Gallatin as secretary of the Treasury, Jefferson placed one of the most capable managers of fiscal policy in the most strategic location. Gallatin, a Swiss-born prodigy with impeccable Republican credentials, dominated the cabinet discussions alongside Madison, the ever-loyal Jefferson disciple who served as secretary of state.

The major achievement of his first term was an act of defiance, though this time it involved defying his own principles. In 1803 Napoleon decided to consolidate his resources for a new round of the conflict with England by selling the vast Louisiana region, which stretched from the Mississippi Valley to the Rocky Mountains. Although the asking price, $15 million, was a stupendous bargain, assuming the cost meant substantially increasing the national debt. More significantly, what became known as the Louisiana Purchase violated Jefferson's constitutional scruples. Indeed, many historians regard it as the boldest executive action in American history. But Jefferson never wavered, reasoning that the opportunity to double the national domain was too good to miss. The American West always triggered Jefferson's most visionary energies, seeing it, as he did, as America's future, the place where the simple republican principles could be constantly renewed. In one fell swoop he removed the threat of a major European power from America's borders and extended the life span of the uncluttered agrarian values he so cherished. Even before news that the purchase was approved reached the United States in July 1803, Jefferson dispatched his private secretary, Meriwether Lewis, to lead an expedition to explore the new acquisition and the lands beyond, all the way to the Pacific.

Louisiana Purchase

If the Louisiana Purchase was the crowning achievement of Jefferson's presidency, it also proved to be the high point from which events moved steadily in the other direction. Although the Federalist Party was dead as a national force, pockets of Federalist opposition still survived, especially in New England. Despite his eloquent testimonials to the need for a free press, Jefferson was outraged by the persistent attacks on his policies and character from those quarters, and he instructed the attorneys general in the recalcitrant states to seek indictments, in clear violation of his principled commitment to freedom of expression. He was equally heavy-handed in his treatment of Aaron Burr, who was tried for treason after leading a mysterious expedition into the American Southwest allegedly designed to detach that region from the United States with Burr crowned as its benevolent dictator. The charges were never proved, but Jefferson demanded Burr's conviction despite the lack of evidence. He was overruled in the end by Chief Justice John Marshall, who sat as the judge in the trial.

But Jefferson's major disappointment had its origins in Europe with the resumption of the Napoleonic Wars, which resulted in naval blockades in the Atlantic and Caribbean that severely curtailed American trade and pressured the U.S. government to take sides in the conflict. Jefferson's response was the Embargo Act (1807), which essentially closed American ports to all foreign imports and American exports. The embargo assumed that the loss of American trade would force England and France to alter their policies, but this fond hope was always an illusion, since the embryonic American economy lacked the size to generate such influence and was itself wrecked by Jefferson's action. Moreover, the enforcement of the Embargo Act required the exercise of precisely those coercive powers by the federal government that Jefferson had previously opposed. By the time he left office in March 1809, Jefferson was a tired and beaten man, anxious to escape the consequences of his futile efforts to preserve American neutrality and eager to embrace the two-term precedent established by Washington.

Embargo Act

Retirement. During the last 17 years of his life Jefferson maintained a crowded and active schedule. He rose with the dawn each day, bathed his feet in cold water, then spent the morning on his correspondence (one year he counted writing 1,268 letters) and working in his garden. Each afternoon he took a two-hour ride around his grounds. Dinner, served in the late afternoon, was usually an occasion to gather his daughter Martha and her 12 children, along with the inevitable visitors. Monticello became a veritable hotel during these years, on occasion housing 50 guests. The lack of privacy caused Jefferson to build a separate house on his Bedford estate about 90 miles (140 km) from Monticello, where he periodically fled for seclusion.

Three architectural projects claimed a considerable share of his attention. Throughout his life Monticello remained a work-in-progress that had the appearance of a construction site. Even during his retirement years, Jefferson's intensive efforts at completing the renovations never quite produced the masterpiece of neoclassical design he wanted to achieve and that modern-day visitors to Monticello find so compelling. A smaller but more architecturally distinctive mansion at Bedford, called Poplar Forest, was completed on schedule. It too embodied neoclassical principles but was shaped as a perfect octagon. Finally there was the campus of the University of Virginia at Charlottesville, which Jefferson called his "academical village." Jefferson surveyed the site, which he could view in the distance from his mountaintop, and chose the Pantheon of Rome as the model for the rotunda, the centrepiece flanked by two rows of living quarters for students and faculty. In 1976 the American Institute of Architects voted it "the proudest achievement of American architecture in the past 200 years."

The University of Virginia

In 1812 his vast correspondence began to include an exchange with his former friend and more recent rival John Adams. The reconciliation between the two patriarchs was arranged by their mutual friend Benjamin Rush, who described them as "the North and South poles of the American Revolution." That description suggested more than merely geographic symbolism, since Adams and Jefferson effectively, even dramatically, embodied the twin impulses of the revolutionary generation. As the "Sage of Monticello," Jefferson represented the Revolution as a clean break with the past, the rejection of all European versions of political discipline as feudal vestiges, the ingrained hostility toward all mechanisms of governmental authority that originated in faraway places. As the "Sage of Quincy (Massachusetts)," Adams resembled an American version of Edmund Burke, which meant that he attributed the success of the American Revolution to its linkage with past practices, most especially the tradition of representative government established in the colonial assemblies. He regarded the constitutional settlement of 1787–88 as a shrewd compromise with the political necessities of a nation-state exercising jurisdiction over an extensive, eventually continental, empire, not as a betrayal of the American Revolution but an evolutionary fulfillment of its promise.

One issue that even Adams and Jefferson could not discuss candidly was slavery. Jefferson's mature position on that forbidden subject represented a further retreat from any leadership role in ending the "peculiar institution." In 1819, during the debate in Congress over the Missouri Compromise, he endorsed the expansion of slavery into all the western territories, precisely the opposite of the position he had taken in the 1780s. Though he continued to insist that slavery was a massive anomaly, he insisted even more strongly that it was wrong for the federal government to attempt any effort at emancipation. In fact he described any federal intrusion in the matter as a despotic act analogous to George III's imperial interference in colonial affairs or Hamilton's corrupt scheme to establish a disguised form of monarchy in the early republic. His letters to fellow Virginians during his last years reflect a conspiratorial mentality toward the national government and a clear preference for secession if threatened with any mandatory plan for abolition.

Apart from slavery, the other shadow that darkened Monticello during Jefferson's twilight years was debt. By the end, he was more than $100,000—in modern terms several million dollars—in debt. Monticello, including land, mansion, furnishings, and the vast bulk of the slave population, was auctioned off the year after his death, and his surviving daughter, Martha, was forced to accept charitable contributions to sustain her family.

Jefferson went to his maker in his bed at Monticello at about half past noon on July 4, 1826. His last conscious words, uttered the preceding evening, were "Is it the Fourth?" Always a man given to Herculean feats of self-control, he somehow managed to time his own death to coincide with history. More remarkably, up in Quincy on that same day his old rival and friend also managed to die on schedule. John Adams passed away later in the afternoon. His last words—"Thomas Jefferson still lives"—were wrong at the moment but right for the future, since Jefferson's complex legacy was destined to become the most resonant and controversial touchstone in all of American history. (Jo.E.)

BIBLIOGRAPHY. More has been written about Thomas Jefferson than any other figure in American history save Abraham Lin-

coln. The authoritative bibliography is FRANK SHUFFELTON, *Thomas Jefferson: A Comprehensive, Annotated Bibliography of Writings About Him (1826–1980)* (1983), continued in *Thomas Jefferson, 1981–1990: An Annotated Bibliography* (1992).

As for Jefferson's own writings, the definitive edition of Jefferson's papers is JULIAN P. BOYD *et al.* (eds.), *Papers of Thomas Jefferson* (1950–), which includes extensive editorial notes, all of Jefferson's known letters and writings, plus all correspondence to him that has survived. This multivolume work is an ongoing project; but this edition stops in the mid-1790s, so material relating to his public career as vice president and president, as well as his long retirement, must be obtained from two older editions: ANDREW ADGATE LIPSCOMB (ed.), *The Writings of Thomas Jefferson*, 20 vol. (1903–04), is more comprehensive but less reliable; PAUL LEICESTER FORD (ed.), *The Works of Thomas Jefferson*, 12 vol. (1893–99), is more reliable but less comprehensive. Convenient collections of essential primary sources are MERRILL D. PETERSON (ed.), *The Portable Thomas Jefferson* (1975, reissued 1997), and *Writings* (1984). Several collections of correspondence provide access to specific phases of his career or particular relationships: JAMES MORTON SMITH (ed.), *The Republic of Letters*, 3 vol. (1995), reproduces the correspondence with James Madison from 1776 to 1826; DOUGLAS L. WILSON and LUCIA STANTON (eds.), *Jefferson Abroad* (1999), covers the years in France, 1784 to 1789; LESTER J. CAPPON (ed.), *The Adams-Jefferson Letters*, 2 vol. (1959, reprinted 1 vol., 1988), provides the extraordinary correspondence with John and Abigail Adams from 1777 to 1826; and EDWIN MORRIS BETTS and JAMES ADAM BEAR, JR. (eds.), *The Family Letters of Thomas Jefferson* (1966, reprinted 1986), contains a good sample of letters to and from his daughters and grandchildren. Information about Jefferson's library, which became the basis for the Library of Congress collection when Jefferson sold it to the federal government in 1815, is available in E. MILLICENT SOWERBY (compiler), *Catalogue of the Library of Thomas Jefferson*, 5 vol. (1952–59, reprinted 1983).

Among the full-life biographies, the standard against which all are measured is DUMAS MALONE, *Jefferson and His Time*, 6 vol. (1948–81), which is a monumental scholarly achievement that takes Jefferson's view of all the controversial issues. Less comprehensive but somewhat more critical, especially on the latter years, is MERRILL D. PETERSON, *Thomas Jefferson and the New Nation* (1970, reprinted 1987). Invaluable because it draws upon conversations with those family and friends who knew Jefferson personally is the early biography by HENRY STEPHENS RANDALL, *The Life of Thomas Jefferson*, 3 vol. (1857, reprinted 1972). Reliable and reverential accounts include GILBERT CHINARD, *Thomas Jefferson: The Apostle of Americanism*, 2nd ed. rev. (1939, reissued 1963); NOBLE E. CUNNINGHAM, JR., *In Pursuit of Reason: The Life of Thomas Jefferson* (1987, reissued 1992); and ALF J. MAPP, JR., *Thomas Jefferson*, 2 vol. (1987–91). FAWN M. BRODIE, *Thomas Jefferson: An Intimate History* (1974, reissued 1998), makes the sexual relationship with Sally Hemings the centre of the story and also makes extensive use of psychiatric theories. ANDREW BURSTEIN, *The Inner Jefferson: Portrait of a Grieving Optimist* (1995), focuses on the sentimentalism of Jefferson's core identity. JOSEPH J. ELLIS, *American Sphinx: The Character of Thomas Jefferson* (1997), provides a more critical portrait, focusing on Jefferson's systemic contradictions.

On Jefferson's political thought, see GARRETT WARD SHELDON, *The Political Philosophy of Thomas Jefferson* (1991, reprinted 1993); DAVID N. MAYER, *The Constitutional Thought of Thomas Jefferson* (1994); and RICHARD K. MATTHEWS, *The Radical Politics of Thomas Jefferson: A Revisionist View* (1984, reissued 1987). For the broader intellectual context, see DANIEL J. BOORSTIN, *The Lost World of Thomas Jefferson* (1948, reissued 1993). Three books offer different interpretations of the Declaration of Independence: CARL BECKER, *The Declaration of Independence: A Study in the History of Political Ideas* (1922, reprinted 1972); GARRY WILLS, *Inventing America: Jefferson's Declaration of Independence* (1978, reissued 1980); and PAULINE MAIER, *American Scripture: Making the Declaration of Independence* (1997). Jefferson's protean political legacy is the subject of MERRILL D. PETERSON, *The Jefferson Image in the American Mind* (1960, reissued 1998), a magisterial account. A major new window into Jefferson's thought is provided by HERBERT E. SLOAN, *Principle and Interest: Thomas Jefferson and the Problem of Debt* (1995).

On slavery and race, JOHN CHESTER MILLER, *The Wolf by the Ears: Thomas Jefferson and Slavery* (1977, reissued 1991), is the standard work. WINTHROP D. JORDAN, *White over Black: American Attitudes Toward the Negro, 1550–1812* (1968, reissued 1977), is seminal; as is DAVID BRION DAVIS, *The Problem of Slavery in the Age of Revolution*, 1770–1823 (1975, reissued 1999). For the context of slavery at Monticello, see JAMES A. BEAR, JR. (ed.), *Jefferson at Monticello* (1967). The great paradox of Jefferson's position as a slaveholder is probed rather brilliantly in EDMUND S. MORGAN, *American Slavery, American Freedom: The Ordeal of Colonial Virginia* (1975, reissued 1995).

The domestic context at Monticello is the subject of several important books. SARAH N. RANDOLPH, *The Domestic Life of Thomas Jefferson* (1871, reprinted 1978), provides an affectionate portrait by his descendants. JACK MCLAUGHLIN, *Jefferson and Monticello: The Biography of a Builder* (1988, reissued 1990), goes beyond architectural issues to explore the daily lives of all the residents on the mountain. ELIZABETH LANGHORNE, *Monticello: A Family Story* (1987, reissued 1989), focuses on the white residents. EDWIN MORRIS BETTS (ed.), *Thomas Jefferson's Farm Book* (1953, reprinted 1987), reproduces Jefferson's plantation records. SUSAN R. STEIN, *The Worlds of Thomas Jefferson at Monticello* (1993), recovers the material objects Jefferson gathered inside the mansion. MERRILL D. PETERSON (ed.), *Visitors to Monticello* (1989), reproduces the accounts of firsthand observers.

Jefferson's five-year stay in France is explored in WILLIAM HOWARD ADAMS, *The Paris Years of Thomas Jefferson* (1997); GEORGE GREEN SHACKELFORD, *Thomas Jefferson's Travels in Europe, 1784–1789* (1995); and MARIE GOEBEL KIMBALL, *Jefferson: The Scene of Europe, 1784 to 1789* (1950). On the matter of Jefferson's love affair with the French Revolution, CONOR CRUISE O'BRIEN, *The Long Affair: Thomas Jefferson and the French Revolution, 1785–1800* (1996, reissued 1998), is a highly critical account.

On his political career in the 1790s, the standard overview is LANCE BANNING, *The Jeffersonian Persuasion: Evolution of a Party Ideology* (1978, reissued 1980). See also JOSEPH CHARLES, *The Origins of the American Party System* (1956, reissued 1961); and NOBLE E. CUNNINGHAM, *The Jeffersonian Republicans: The Formation of Party Organization, 1789–1801* (1957, reissued 1963). The crucial role of the political partnership with James Madison is addressed in ADRIENNE KOCH, *Jefferson and Madison: The Great Collaboration* (1950, reprinted 1986).

In addition to the coverage provided in the full-scale biographies, Jefferson's presidency receives book-length treatment in FORREST MCDONALD, *The Presidency of Thomas Jefferson* (1976); ROBERT M. JOHNSTONE, JR., *Jefferson and the Presidency: Leadership in the Young Republic* (1978); and NOBLE E. CUNNINGHAM, JR., *The Process of Government Under Jefferson* (1978). Foreign policy is best discussed in ROBERT W. TUCKER and DAVID C. HENDRICKSON, *Empire of Liberty: The Statecraft of Thomas Jefferson* (1990, reissued 1992); LAWRENCE S. KAPLAN, *Entangling Alliances with None: American Foreign Policy in the Age of Jefferson* (1987); and GEORGE DARGO, *Jefferson's Louisiana: Politics and the Clash of Legal Traditions* (1975). For the domestic implications of the embargo crisis, a splendid assessment is LEONARD WILLIAMS LEVY, *Jefferson & Civil Liberties: The Darker Side* (1963, reprinted 1989). On his battles with the judiciary, see RICHARD E. ELLIS, *The Jeffersonian Crisis: Courts and Politics in the Young Republic* (1971, reprinted 1974).

Each of Jefferson's multiple dimensions has attracted scholarly interest. For his scientific achievements, see SILVIO A. BEDINI, *Thomas Jefferson: Statesman of Science* (1990); and I. BERNARD COHEN, *Science and the Founding Fathers* (1995, reissued 1997). For his religious views, EDWIN S. GAUSTAD, *Sworn on the Altar of God: A Religious Biography of Thomas Jefferson* (1996). For his aesthetic and architectural contribution, WILLIAM HOWARD ADAMS (ed.), *Jefferson and the Arts: An Extended View* (1976). For his educational vision, ROY J. HONEYWELL, *The Educational Work of Thomas Jefferson* (1931, reissued 1964); and HAROLD HELLENBRAND, *The Unfinished Revolution: Education and Politics in the Thought of Thomas Jefferson* (1990).

On the question of Jefferson's relationship with Sally Hemings, the pre-DNA argument against the liaison is best summarized in VIRGINIUS DABNEY, *The Jefferson Scandals: A Rebuttal* (1981, reissued 1991). The argument for the liaison is nicely synthesized by ANNETTE GORDON-REED, *Thomas Jefferson and Sally Hemings: An American Controversy* (1997). The DNA study appears in EUGENE A. FOSTER *et al.*, "Jefferson Fathered Slave's Last Child," *Nature*, 396(6706):27–28 (November 5, 1998).

(Jo.E.)

Jerusalem

Jerusalem (Hebrew Yerushalayim; Arabic Bayt al-Muqaddas, or al-Quds), one of the world's oldest and holiest cities, was in December 1949 proclaimed by the State of Israel to be its capital. The city plays a central role in the spiritual and emotional perspective of the three major monotheistic religions. For Jews throughout the world, Jerusalem is the focus of age-old yearnings, a living proof of ancient grandeur and independence and a centre of national renaissance; for Christians, it is the scene of their Saviour's agony and triumph; for Muslims, it is the goal of the Prophet Muḥammad's mystic night journey and the site of one of Islām's most sacred shrines. For all three faiths it is a centre of pilgrimage—the Holy City, the earthly prototype of the heavenly Jerusalem.

Following the collapse of the British mandate in Palestine after World War II, the United Nations (UN) declared Jerusalem an international city. The failure of this endeavour and the outbreak of the first Arab-Israeli war in 1948 left Jerusalem divided into Israeli (West Jerusalem) and Jordanian (East Jerusalem) sectors, with West Jerusalem as Israel's capital. During the Six Day War of 1967, Israel occupied the Jordanian sector and, despite international protest, annexed East Jerusalem and parts of the occupied West Bank, proclaiming the combined city its capital.

In subsequent decades this annexation proved to be a point of great contention between Israel and the Arab and international communities. Despite numerous UN resolutions calling for Israel's withdrawal, the Jewish state continued its policy of constructing satellite settlements around Jerusalem—both within its self-declared city limits and outside those boundaries in adjacent sections of the West Bank—and in 1980 reiterated its stance by passing a quasi-constitutional "basic law" mandating a "complete and united" Jerusalem as Israel's capital.

The Palestinian Arabs of Jerusalem, the West Bank, and Gaza continue to decry these annexations and, in turn, insist that East Jerusalem be the capital of any Palestinian state. The status of the city has become a central issue in the dispute between the two sides.

Despite a rapidly changing demography and an increasing number of Jewish residents, Jerusalem, which covers an area of 42 square miles (109 square kilometres), has maintained a diverse and cosmopolitan character, particularly the Old City with its Armenian, Christian, Jewish, and Muslim quarters. Arabs in traditional and modern dress; Christians, Western and Oriental, in their infinite variety of secular and monastic vestments; Jews in fashionable and Orthodox dress; and hosts of tourists combine in colourful, kaleidoscopic patterns. Synagogues, churches, mosques, and dwellings in various styles make up the city's unique architectural mosaic. The scent of Oriental cooking and spices, the peal of church bells, the calls of muezzins from minarets, and the chanting of Jewish prayers at the Western (Wailing) Wall all add a particular tinge to the life of the city. These impressions, however, are in a large measure limited to the Old City. Outside the walls Jerusalem is in every sense a modern city with its network of streets and transportation, high-rise buildings, supermarkets, businesses, schools, and restaurants and coffeehouses. It is the persistent mingling of Hebrew, Arabic, and English in the streets that brings to mind the multicultural and political complexities of life in this revered city.

This article is divided into the following sections:

Physical and human geography

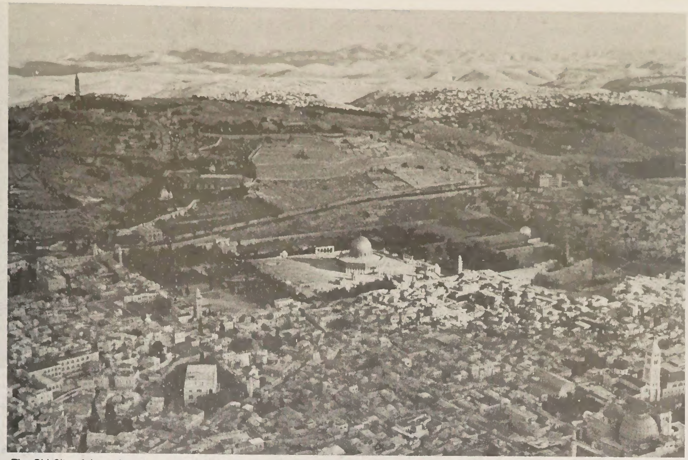
THE LANDSCAPE

The city site. To the east, Jerusalem looks down on the Dead Sea and across the Jordan River to the arid mountains of Jordan; to the west, it faces the coastal plain and the Mediterranean Sea, about 35 miles (58 kilometres) away. The main north-south road bisects the city in its course along the watershed between the coastal plain and the Great Rift Valley of the Jordan River and links Nābulus to the north with Bethlehem, Hebron, and Beersheba to the south. A major road links Jerusalem with Jericho, about 36 road miles to the east, and hence along the Jordan to the Sea of Galilee in the north. The (Yigal) Allon Road cuts across the Judaean Desert, linking with the new Israeli settlements in the West Bank. The west-east road from Tel Aviv–Yafo, 58 miles to the west, crosses the Jordan north of the Dead Sea and runs to Amman, Jordan, about 60 miles to the east; the Tel Aviv–Jerusalem sector has been developed into a modern four-lane highway. A newer transversal road leads westward from Jerusalem, eventually converging on the Ben-Gurion Airport at Lod.

Climate. Jerusalem has a mixed subtropical, semiarid climate with warm, dry summers and cool, rainy winters. The average annual precipitation is about 20 inches (500 millimetres), and snow falls every two or three years. Average temperatures range from about 75° F (24° C) in August to about 50° F (10° C) in January. The hot desert wind, called *sharav* (*khamsīn*), is fairly common in autumn and spring. Average daily humidity is about 62 percent in the daytime but may drop 30–40 percent under *sharav* conditions. Summer exposure to the Sun's rays in Jerusalem is among the most intense on the globe, attributable partly to the lack of clouds or humidity and partly to the angle of the Sun (80°) over the horizon at that season.

Jerusalem has no serious air pollution. Its elevation ensures the free mixing of surface air, and pollutant sources are few, for there is little heavy industry.

Plant and animal life. Lying on the watershed between the relatively rainy Har Yehuda ("Hills of Judaea") and the dry Judaean Desert, Jerusalem has both Mediterranean and Irano-Turanian vegetation. The various red and brown Mediterranean soils, formed by the different types of limestone chalk covering the hills, support as many as 1,000 plant species.



The Old City of Jerusalem. In the centre is the Dome of the Rock; the Dead Sea is in the background.
Werner Braun

There is a great variety of birds, including 70 resident species and about 150 winter visitors. Those most commonly seen are the hooded crow, jay, swift (which nests in old walls and buildings), and bulbul. The only venomous snake is the Palestine viper; the smooth lizard and common chameleon frequent gardens.

The city layout. The disputed municipal boundaries, which were defined in 1967, stretch from the Jerusalem Airport in the north to a point almost reaching Bethlehem in the south and from the ridge of Mount Scopus and the Mount of Olives in the east to Mount Herzl, ʿEn Kerem, and the Hadassah Medical Centre of Hebrew University in the west.

The Old City

The Old City, which is believed to have been continuously inhabited for almost 5,000 years, forms a walled quadrilateral about 3,000 feet (900 metres) long on each side. It is dominated by the raised platform of the Temple Mount (Hebrew: Har ha-Bayt; Arabic: al-Ḥarām ash-Sharīf, "The Noble Sanctuary"), commonly held to be the site of the First and Second Temples and upon which reside numerous structures venerated by Muslims. The rest of the area within the walls is a typical Middle Eastern city, with its mosques and medieval vaulted triple bazaars in the centre and a labyrinth of smaller *suqs,* or bazaars, along David Street, which leads from Jaffa Gate and the Citadel toward the Temple Mount. The Old City is distinguished by its many churches and by the ancient synagogues and study houses of the Jewish Quarter.

Neighbourhoods outside the walls

The first neighbourhoods outside the Old City walls, built from the 1860s onward, were scattered chiefly along the main roads leading into the city. The earliest of the Jewish communities were paralleled by non-Jewish expansion prompted by Christian religious or nationalistic motivation and included establishment of the Russian Compound, the German Colony, and the American Colony. Some early communities, such as Mishkenot Shaʾanannim and Yemin Moshe, have been reconstructed and partially settled or turned into cultural centres. Others include the Bukharan Quarter; Meʾa Sheʿarim, founded by Orthodox Jews from eastern and central Europe, with its scores of small synagogues and Talmudic study houses; and Maḥane Yehuda, with its fruit and vegetable market, inhabited mainly by Sephardic Jews. Residential quarters established between World Wars I and II include Reḥavya in the centre, Talpiyyot in the south, and Qiryat Moshe and Bet ha-Kerem in the west. The old campus of the Hebrew University at Mount Scopus, which formed for 20 years (1948–67) an Israeli enclave in the Jordanian-dominated sector, was entirely rebuilt after the Six-Day War. Some Arab districts, such as Talbieh (modern Qomemiyyut) and Katamon (Gonen), whose residents fled during the fighting of 1947–48, were occupied by Jews, and thousands of houses were built for new immigrants in districts to the west, newly incorporated into the city. Arab neighbourhoods outside the Old City include the American Colony, ash-Shaykh Jarrāḥ, Wādī al-Jōz, and Bayt Ḥanīnā in the north and villages such as Silwān and Bayt Ṣafāfā in the south. Jewish settlements in areas surrounding East Jerusalem include Gillo, Newe Yaʿaqov, and Ramot Allon.

Housing. There is a great variety of housing in the city. In the Old City are antiquated buildings constructed of ancient stones; 19th-century Jewish neighbourhoods, some of which have declined into slums; modern quarters with tree-lined streets; and government-built housing projects, mainly for new immigrants. The most common basic dwelling unit in the Old City consists of a complex of structures, often on different levels, built around an inner court that is entered through a narrow corridor.

Architecture. The outstanding characteristic of the architecture of Jerusalem is the coexistence of old and new, sacred and secular, in a variety of styles. The most conspicuous feature is the Old City Wall, erected 1538–40 by the Ottoman sultan Süleyman the Magnificent, largely on the foundations of earlier walls dating chiefly to the period of the Crusades but in some places dating to Byzantine, Herodian, and even Hasmonean times. On three sides of the Temple Mount, parts of the original supporting walls still stand. During the centuries when Jews were excluded from the Temple Mount, its Western Wall became Jewry's holiest shrine. Since 1967 the wall has been further exposed, and plans have been made to landscape the area once excavations are completed. The main buildings on the platform are the two Islāmic structures: the gold-capped Dome of the Rock, completed in 691, and the silver-domed al-Aqṣā Mosque, built in the early 8th century.

The Wall of Süleyman

The Old City of Jerusalem.

The Citadel (with David's Tower) beside the Jaffa Gate, which acquired its present form in the 16th century, was created over ruins from the Hasmonean and Herodian periods, integrating large parts of crusader structures and some Mamlūk additions. The large number of churches mainly represent two great periods of Christian architecture, the Byzantine and crusader periods. The predominant characteristic of the former are monumental, two- or three-tiered ornamental or carved basketlike capitals, the latter reflecting Romanesque styling, which features pointed arches and ribbed vaults. The Church of the Holy Sepulchre incorporates elements of both styles, but its facade and layout are architecturally Romanesque. The best example of the mixed style is the Church of St. Anne (its substructure is Byzantine); others are the Armenian Cathedral of St. James, which combines Romanesque with Oriental elements, and the Tomb of the Virgin, which is Romanesque in its upper part but Byzantine in its lower. The central part of the triple bazaar, as well as its link with the Cardo (a restored Roman-Byzantine mall), is of crusader origin. Mamlūk constructions of the 13th to the 15th century, as well as coats of arms of Mamlūk rulers, are found along David Street and near the Gate of the Chain at the Western Wall. The constructions are characterized by "stalactite" or "honeycomb" ornamentation and the use of multicoloured slabs of stone. Ottoman architecture of the early 16th century continued the Mamlūk style and is represented in some structures of the Temple Mount. The rock-cut tombs east and north of the Old City exemplify architecture of the first half of the 1st millennium BC (Tomb of Pharaoh's Daughter) and the Second Temple period (Tombs of the Kings, Tomb of Absalom, and Tomb of Zechariah). The restored Monastery of the Cross, in the heart of modern Jerusalem, dates from the 5th century.

As Jerusalem spread outside the walls, the architecture came to be characterized chiefly by iron beams and red-tiled roofs. From 1930 there was a radical change, and flat roofs and reinforced concrete faced with naturally dressed

Jerusalem.

stone predominated. Whereas residential buildings are seldom taller than four stories and office buildings seldom taller than eight, there is a growing tendency, despite opposition, for high-rise construction. This is the case with a number of modern hotels at the western entrance to the city, and the construction of office buildings in the city centre is following the trend. All building, however, must follow a city ordinance requiring construction of stone. Outstanding modern architecture is reflected by the buildings on the university campuses on Mount Scopus and Giv'at Ram, the Knesset (Parliament), the Israel Museum, the Jerusalem Theatre, and the Hebrew Union College. More modern trends are represented by the Bank of Israel, the Jerusalem Great Synagogue, and the President's Home. An earlier generation is represented by the Government House (UN Headquarters), the King David Hotel, the Rockefeller (Archaeological) Museum, and the Young Men's Christian Association.

THE PEOPLE

Because Jerusalem is a holy city, uniquely revered by three major religions, its people can, perhaps, be best described

Composition of the population

according to religious affiliation. Most of the city's residents are Jews. The Muslims are the most homogeneous of the communities, the Christians the most diversified.

Responsibility for the city's holy places and religious communities is vested in the Ministry of Religious Affairs, which has special desks for the individual denominations. The administration, protection, and care of holy places are in the hands of the respective religious authorities. Penalties of up to seven years' imprisonment may be inflicted for desecration of these places.

Despite this, a small but vocal minority among the Jewish community has advocated the removal of Islāmic structures from the Temple Mount and the erection of a Third Temple on the site. Although Israel's official policy strictly abjures such actions, there have been several attempts to occupy or damage Islāmic sites on the Mount. These actions have strained relations between Jews and minority religious groups, particularly the Muslims, who feel that the Israeli government favours Jewish extremists in such confrontations.

Jews. Among the Jews, the main divisions are between Ashkenazim and Sephardim, names denoting, somewhat inaccurately, places of origin. Of more importance is the division between the Orthodox and the more secular-minded segments of the population, whose attitudes over religious as well as political matters often conflict. Jerusalem is the centre of Jewish religious reverence and aspiration. The most sacred spot is the Temple Mount, on which many Orthodox Jews refrain from setting foot for fear of profaning its sanctity. In addition to the Western Wall—the most important centre of prayer and pilgrimage—other holy places are Mount Zion, with the reputed tomb of King David, the Mount of Olives, with its ancient Jewish cemetery, and the tombs of priestly families in the Valley of Kidron. Ancient synagogues and study houses in the Old City are being restored; particularly worthy of mention is the interconnected group of four synagogues begun in the 16th century by Jews exiled from Spain.

Muslims. Jerusalem's Muslims are mostly Sunnites, though the city is revered by all Muslims as the third holiest place on earth, and the pilgrimage to Jerusalem (*taqdīs*) completes the main pilgrimage (*ḥajj*). Muslim religious sites, including those on the Temple Mount, have traditionally been funded and administered through the Islāmic *waqf* (religious endowment) under the control of Jordan's Council of Waqf and Muslim Affairs. Although Jordan surrendered control of *waqf* property in the West Bank to the Palestinian Authority (PA) in the mid-1990s, the Hāshimite monarchy has sought to maintain its position as protector of the holy places and has continued to administer the *waqf* in East Jerusalem and appoint the city's supreme Muslim figure, the mufti. This has led to a conflict with the PA, which has sought control of the *waqf*, has appointed its own mufti, and desires to obtain control of East Jerusalem (and its holy places) as the capital of a Palestinian state.

Christians. Christians constitute the smallest and most highly diversified section of the population. The city is the seat of three resident patriarchs of the Eastern Orthodox churches and many archbishops and bishops, and it has an ecclesiastical embassy for almost every sect in Christendom. The main groups are Eastern Orthodox, Monophysite, Roman Catholic, and Protestant. Major denominations share control over the Church of the Holy Sepulchre. Most of the church bodies in Jerusalem maintain scholarly research institutes with fine libraries and archives.

The Greek Orthodox Church maintains a patriarchate with jurisdiction over the entire Holy Land. The Russian Orthodox churches (one governed from Moscow and one in exile) have considerable properties dating to tsarist times. The Roman Catholic Church in Jerusalem, established in 1099 during the First Crusade, was dissolved when the Muslims won the city in 1244. The Latin Patriarchate was reestablished in 1847. Of the Monophysite churches, the Armenian is the largest (others include the Coptic and the Abyssinian), its patriarchate having been established in the 6th century. The Protestant community is small but influential.

Armenian Church

THE ECONOMY

The main source of livelihood in Jerusalem is government and public service employment (including the academic and clerical professions). Since 1967, business activity and investment in the city have been stimulated by the housing boom and the ever-increasing influx of pilgrims and tourists. Personal income has risen steadily, and unemployment is marginal, although the city still deals with a large number of social welfare cases.

Industry and trade. The establishment of heavy industries has not been encouraged, in the interest of preserving the traditional character of the city. Combined with transport and marketing difficulties, this has limited the city to a number of small industries. They include diamond cutting and polishing and the manufacture of home appliances, furniture, shoes, pencils, plastics, textiles, clothing, and pharmaceuticals and chemicals. There are also printing and publishing houses, as well as workshops producing jewelry, giftware, religious articles, curios, and printed fabrics. More recent additions include the modern science-based industries and the development of industrial quarters in the outskirts of the city and in some of its easily accessible satellite settlements in Israel and the West Bank. Nevertheless, the percentage of the workforce engaged in industry remains quite small, whereas about two-thirds is engaged in services. The tourist boom has stimulated the construction of first-rate hotels in the city, which receives the highest number of tourists in the country. The heaviest influx is linked with the Jewish high holidays, Christmas, Passover, Easter, and the Muslim pilgrimage.

Transportation. Despite a considerable increase in road construction since the mid-1970s, traffic congestion remains one of the most acute problems in city planning because of the ever-increasing number of private vehicles. Public transportation is provided by bus companies, which operate in eastern and western Jerusalem and also make interurban connections. The latter are also offered by shuttle services (*sherut*), which connect Jerusalem with all major Israeli towns, including Elat and Tiberias. The old railway connecting Jerusalem with Tel Aviv–Yafo and Haifa on the coast and with Beersheeba inland is of secondary importance; the Jerusalem Airport at the northern edge of the city serves mainly inland tourist traffic.

ADMINISTRATION AND SOCIAL CONDITIONS

Government. Jerusalem is the seat of the president and the Knesset (Parliament) of Israel. Although a large number of countries did not—and still do not—recognize Jerusalem as the capital of Israel, a large proportion of the resident foreign embassies and legations in Israel were located in Jerusalem until the 1980 basic law officially proclaimed the unified city the nation's capital. At UN urging, most of the foreign delegations then moved to Tel Aviv–Yafo to reaffirm nonrecognition of Jerusalem as the Israeli capital. France and the United States each maintain consulates in the eastern and western parts of the city. Diplomats living in the Tel Aviv area go to Jerusalem to present their credentials to the president and transact business at the Foreign Ministry. The ministries are concentrated in the Qiryat Ben-Gurion, the government complex, which is flanked by the Knesset on one side and the Bank of Israel on the other. The Ministry of Defense is still located in Tel Aviv–Yafo, and several ministries are in temporary housing in Jerusalem. In addition to the Supreme Court and the Chief Rabbinate, the city also houses the head offices of many world Jewish bodies, such as the Jewish Agency and the World Zionist Organization, as well as the Martyrs' and Heroes' Remembrance Authority (Yad va-Shem), which commemorates the victims of the Holocaust.

The Municipal Council is composed of 31 members who are elected every four years. The council is headed by the mayor, who, since 1975, has been elected by direct vote. Permanent residents, even if not Israeli nationals, are entitled to vote. Official correspondence is issued in both Hebrew and Arabic.

Voting

Services. Jerusalem has always depended on human ingenuity to bring its water from afar. The underground aqueduct thought to have been built by King Hezekiah in the 8th century BC is still extant, and many reservoirs and rainwater cisterns date from ancient times. Since the 1950s the New City has enjoyed a limited supply from the Israeli national water grid; East Jerusalem was connected to the West Jerusalem system in 1967.

The city has a modern sewerage system. The six miles of ancient piping that run through the Old City still present serious engineering problems. Drainage repairs in the Christian Quarter have uncovered Byzantine pavements, which have now been restored. Additionally, parts of the Via Dolorosa, said to follow the path along which Jesus carried the cross to Golgotha, have been repaved to facilitate the Christian Holy Week pilgrimage.

Electricity is supplied by the national grid of an Israeli government corporation, as well as by a small diesel plant in East Jerusalem.

Health. The Hadassah Medical Centre at ʿEn Kerem, one of the most advanced institutions of its kind in the world, treats patients from throughout the country, as well as from the West Bank, the Gaza Strip, and Jordan. Other hospitals are the Hadassah Hospital on Mount Scopus; Shaʿare Tzedeq, which pays special attention to the requirements of Orthodox Jews; Biqur Ḥolim; St. John's Ophthalmic Hospital; Ezrat Nashim for mental patients; Alyn for handicapped children; an Arab-Muslim hospital, al-Maqāṣid al-Khayrīyah, at et-Tur; and an Arab-Christian hospital, al-Muṭallaʿ (Augusta Victoria), on the Mount of Olives, which is run by Lutheran organizations that mainly care for the Arab population. A modern medical centre that also serves the Arab population was opened in 1982 at ash-Shaykh Jarrāḥ in northeast Jerusalem. Also important are the Austrian Hospice inside the old town, the French Hospital, St. Louis (accepting terminal cases), and the Sisters of Charity (for the handicapped). After unification of East and West Jerusalem the Kupat Ḥolim, the medical insurance arm of the General Federation of Labour, established several clinics in the eastern part of the city. Supplementing the regular medical facilities are the Magen David Adom and the Red Crescent (counterparts of the Red Cross), which provide additional emergency and medical services.

Religious and secular education

Education. The language of instruction is Hebrew in Jewish schools and Arabic in Arab schools. Hebrew and Arabic are alternately first or second languages in all schools. In structure and curriculum, as a rule, the Arab schools follow the Jordanian system. While the majority of school-age children attend government schools, there are numerous private institutions maintained by Jewish, Muslim, and Christian religious organizations. In the latter, the language of instruction is sometimes French or English. State kindergartens were introduced in East Jerusalem in 1967. Education is the single most important item in the city's budget, and the municipality is responsible for the maintenance of classrooms from kindergarten through high school.

The Hebrew University of Jerusalem (opened 1925) is Israel's foremost institution of higher learning. It has two main campuses—at Mount Scopus in the east and at Givʿat Ram in the west, in addition to the medical school at ʿEn Kerem and the Faculty of Agriculture in Reḥovot. The old buildings on Mount Scopus have been renovated and supplemented by a new complex of buildings. Other institutes of higher learning are the Bezalel Academy of Arts and Design, the Rubin Academy of Music, the Hebrew Union College, several teacher-training colleges, and an Armenian seminary. Al-Quds University, with its main campus in Abū Dīs in the West Bank, is the only Palestinian institution to offer classes in Jerusalem at the postsecondary level.

Libraries

The Jewish National and University Library, with some 4,000,000 volumes, is the largest in the country. It has the foremost collection of books, incunabula, and periodicals of Judaica in the world, in addition to an excellent library on archaeology and Oriental studies, including the history of Palestine. In addition there are the Library of the Knesset and the State Archives (each of which receives a copy of every book printed in Israel) and the Municipal Library and its branches. Numerous other libraries serve a variety of needs.

CULTURAL LIFE

An important cultural institution is the Israel Museum, which, in addition to its general art collection, houses a comprehensive Middle Eastern archaeological collection, several important Dead Sea Scrolls and other relics, a notable collection of Jewish ritual art, Middle Eastern ethnological exhibits, a sculpture garden, and a youth wing. The Rockefeller Museum concentrates on the archaeology of the Holy Land, and there is an Islāmic Museum near the al-Aqṣā Mosque, as well as the L.A. Mayer Memorial Institute for Islāmic Art in West Jerusalem.

Special mention should be made of the École Biblique et Archéologique Française, the Studium Biblicum Franciscanum, the Pontifical Biblical Institute, the British School of Archaeology, the William Foxwell Albright Institute of Archaeological Research, the Swedish Theological Institute, and the Ben Zvi Institute. All of these have libraries dealing with theology and the ancient and modern history of Israel and the Middle East; some have collections of antiquities and valuable manuscripts. In addition, the Institute for Palestine Studies, the world's premier body for Palestinian research, maintains an office in the city, and the Arab Studies Society has an extensive archive and research centre at Orient House for the study of Arab and Palestinian culture.

Art exhibitions are held in the Israel Museum, the Artists' House, Hutzot ha-Yotzer (the craftsmen's centre), the International Cultural Centre for Youth, and private galleries. Concerts and theatre performances are given at Binyane ha-ʾUma (the Convention Centre), the Khan (housed in a restored 18th-century building), and the Wise Auditorium of the Hebrew University. The beginnings of an Arab theatre have been established.

The only English-language daily in Israel, the *Jerusalem Post,* is printed in Jerusalem, as are all of the Arabic-language dailies. Most Hebrew scholarly periodicals are also printed in the city. The Government Press Office is located at Beit Agron, the headquarters of the Jerusalem Journalists' Association. The headquarters of the Israel Broadcasting Authority (television and radio) are also in Jerusalem. Radio broadcasts are mainly in Hebrew and Arabic, though some programs are broadcast in such languages as English, French, Ladino, Romanian, Russian, and Yiddish.

The Jerusalem Foundation, established in 1966, recruits funds for the preservation of the city's multi-religious heritage and the embellishment of its barren areas. This foundation is responsible for Jerusalem's many parks, gardens, woodlands, and forests. The largest is Jerusalem Park, designed as a greenbelt to encircle the Old City walls. There are also small gardens, playgrounds, and recreational areas dotting the city. The Biblical Zoo houses specimens of all the animals that are mentioned in the Bible, and the Natural History Museum focuses on the country's fauna.

The municipality, the Young Men's Christian Association (YMCA), and local clubs run comprehensive sports programs. The YMCA association football (soccer) field can accommodate 10,000 spectators, and the ha-Poʿel (Workers' Sports Club) field 7,500. There are a number of open-air swimming pools. Community centres in the suburbs also provide sports facilities.

Archaeological excavations

Since 1968 extensive excavations have been carried out in the Old City on behalf of the Hebrew University Institute of Archaeology, the Israel Department of Antiquities and Museums, and the Israel Exploration Society. The digs around the southern and western walls of the Temple Mount, going down to the Herodian pavements, have revealed the steps leading to the Temple, the priests' underground entrance to the Temple, and many religious objects. There are also notable remains of public buildings alongside a main street. Remains found within the precincts of the First Wall in the Jewish Quarter bore the imprint of burning and destruction during the sack of the city by the Romans in AD 70. For the first time were found walls of structures dating to the 8th and 7th centuries BC. One of these has been identified as the "Broad Wall" described by Nehemiah. A crucified body, from Roman times, with a nail still lodged in the ankle, was discovered in a Jewish tomb at Givʿat ha-Mivtar. Extensive excavations in the Citadel uncovered structures of the Hasmonean, Herodian, crusader, and Mamlūk periods.

Below the Temple Mount outside the walls, impressive remains of an Umayyad palace have been found. The excavations since 1978 in the Mount Ophel and City of David area have revealed Canaanite and early Hebrew settlements, the latter with a wealth of seals, bulls, epigraphical material, and everyday utensils. A most significant discovery was that of the Roman and Byzantine Cardo, running from the vicinity of the Zion Gate through the restored Jewish Quarter to its crusader part and crossing with the Old City bazaars. The street was reconstructed using the ancient pavement, columns, and capitals. The discovery of the Crusaders' Church, hospice, and hospital of the crusader Teutonic Order (12th century AD) in the Jewish quarter and the huge expanse of wall and towers (from the crusader and Ayyūbid periods of the 12th and 13th centuries AD) between the Dung Gate and the Zion Gate is a major contribution to the history of the city. (Jo.Pr./Ed.)

The flurry of archaeological investigation in Jerusalem has not been without political controversy, however. In 1997 the opening of an archaeological tunnel exit along the Western Wall ignited Muslim fears that the excavations might undermine the Islāmic structures on the Temple Mount, and rioting ensued. Likewise, some Jews have contended that renovations and excavations on the Temple Mount begun by the Muslim *waqf* in the late 1990s were intended to destroy Jewish cultural treasures.

History

THE EARLY PERIOD

Ancient origins. The earliest traces of human settlement in the city area, found on a hill to the southeast, are from the late Chalcolithic Period (Copper-Stone Age) and Early Bronze Age (*c.* 3000 BC). Excavations have shown that a settlement existed south of the Temple Mount, and a massive town wall was found just above the Gihon Spring, which determined the location of the ancient settlement. The name, known in its earliest form as Urusalim, is probably of western Semitic origin and apparently means "Foundation of Shalem" ("Foundation of God"). The city and its earliest rulers, the Egyptians, are mentioned in the Egyptian Execration Texts (*c.* 1900–1800 BC) and again in the 14th-century Tell el-Amarna correspondence, which contains a message from the city's ruler, Abdi-Kheba (Abdu-Ḥeba), requiring his sovereign's help against the invading Hapiru (Habiru, ʿApiru). A biblical narrative mentions the meeting of Canaanite Melchizedek, said to be king of Salem (Jerusalem), with the Hebrew patriarch Abraham, and in a later episode it mentions another king, Adonizedek, who headed an Amorite coalition and was vanquished by Joshua.

According to biblical accounts, about the year 1000 BC Jerusalem, on the frontier of Benjamin and Judah, inhabited by a mixed population described as Jebusites, was captured by David, founder of the joint kingdom of Israel and Judah, and the city became the Jewish kingdom's capital. His successor, King Solomon, extended the city and built his Temple on the threshing floor of Araunah (Ornan) the Jebusite. Thus Jerusalem became the place of the royal palace and the sacred site of a monotheistic religion.

Kings of Judah

On Solomon's death the northern tribes seceded. In 922 BC the Egyptian pharaoh Sheshonk I sacked the city, to be followed by the Philistines and Arabians in 850 and Joash of Israel in 786. After Hezekiah became king of Judah, he built new fortifications and an underground tunnel, which brought water from Gihon Spring to the Pool of Siloam inside the city, but he succumbed to the might of Sennacherib of Assyria, who in 701 forced payment of a heavy tribute. In 612 Assyria yielded its primacy to Babylon. Eight years later Jerusalem was despoiled, and its king was deported to Babylon. In 587/586 BC the city and Temple were completely destroyed by Nebuchadrezzar, and the captivity began. It ended in 538 BC when Cyrus II the Great of Persia, who had overcome Babylon, permitted the Jews, led by Zerubbabel, of the Davidic house, to return to Jerusalem. The Temple was restored (515 BC) despite Samaritan opposition, and the city became the centre of the new statehood and its position strengthened when Nehemiah (*c.* 444) restored its fortifications.

Hellenistic and Hasmonean periods. With the coming of Alexander the Great and his victory at Issus in 333 BC, Jerusalem came under Greek influence. After Alexander's death, Palestine fell to the share of his marshal, Ptolemy I Soter, son of Lagus, who had occupied Egypt and had made Alexandria his capital. In the year 198 BC Jerusalem was acquired by the northern dynasty, descended from Seleucus I Nicator, another of Alexander's marshals, which ruled from Antioch (contemporary Antakya, Tur.). The growth of Greek, or pagan, influence affronted the orthodox, whose hostility burst into armed rebellion in 167 BC after the Seleucid Antiochus IV Epiphanes had deliberately desecrated the Temple. The revolt was led by a pious countryman called Mattathias, son of Hasmoneus (Hasmon), and was carried on by his son Judas, known as the Maccabee. The Hasmoneans succeeded in expelling the Seleucids, and Jerusalem regained its position as the capital of an independent state ruled by the Hasmonean dynasty.

Roman rule. Rome had for some time been expanding its authority in Asia, and in 63 BC Pompey captured Jerusalem. A clash with Jewish nationalism was averted for some time by the political skill of a remarkable family whose most illustrious member was Herod the Great. Herod was of Edomite descent, though of Jewish faith, and was allied through his mother with the nobility of Nabataean Petra, the rich Arab state that lay to the east of the Jordan River. In 40 BC Herod, who had distinguished himself as governor of Galilee, was appointed "client" king of Judaea by the Roman Senate. He was the friend of Mark Antony and, after the defeat of Antony by Octavian at Actium in 31 BC, of Octavian himself.

The kingdom of Herod

Herod was king for the next 36 years, during which period Jerusalem reached its peak of greatness, growing in wealth and expanding even beyond the new double line of walls. The Temple Mount esplanade was enlarged with supporting walls (including the Western Wall) to house Herod's greatest creation, the splendid new Temple, which took more than a generation to build. The new royal palace was strengthened by immense towers that were integrated in the older Hasmonean walls, whereas the Temple was defended by a new citadel. An amphitheatre added to the Hellenistic character of the city. Centre of religion, goal of obligatory pilgrimage, and the seat of the ruler and of the autonomous court of the Sanhedrin (Jewish Council of Elders), Jerusalem became a great metropolis of the Hellenistic world. Herod died in 4 BC and was succeeded by his son Archelaus, who was subsequently deposed by the Romans in AD 6 and replaced by the first of a series of Roman procurators. It was under the fifth procurator, Pontius Pilate, that Jesus of Nazareth was put to death.

From AD 41 to 44 the kingdom of Herod was reconstituted for his grandson Herod Agrippa I, upon whose premature death the procurators returned. In 66 the Jews rebelled against Rome, and in 70 the city was besieged and almost wholly destroyed by the Roman forces under Titus. The Temple, Herod's greatest creation, was reduced to ashes. By 130 the city had been partially repopulated, and the Jews again revolted unsuccessfully against Rome from 132 to 135. Hadrian decided to plant a Roman city, Aelia Capitolina, on the site, and the general layout of his town has lasted into the 20th century.

Christian pilgrims early found their way to Jerusalem. It was, however, the conversion to Christianity of Constantine the Great and the famous pilgrimage (326) of his mother, Empress Helena, who found "the True Cross," that made possible the building of the famous shrines in Jerusalem, including the Church of the Holy Sepulchre, and inaugurated one of the city's most splendid and prosperous epochs. The Christian glorification was carried on into the 6th century when, under the emperor Justinian, the Church of Resurrection was rebuilt and many other churches, as well as monasteries and hospices, were established. In 614 this golden age was brought to an end by the Persian invasion, in which the inhabitants of Jerusalem were massacred and the churches destroyed.

The Islāmic and crusader periods. In 638 the Muslim caliph ʿUmar I entered Jerusalem and, according to Muslim historians, discovered the Temple Mount in utter decay and disrepair. ʿUmar immediately set about repairing the site, and in 688–691 the fifth Umayyad caliph, ʿAbd al-Malik ibn Marwān, built the Dome of the Rock. The city, however, lost some of its earlier importance, despite being proclaimed a goal of Muslim pilgrimage, when the caliphate was moved from Damascus to Baghdad. Jerusalem shrank in size, and the new line of walls (11th century) did not include the City of David and Mount Zion. Both the Umayyads and their successors, the ʿAbbāsids, pursued a liberal policy toward Christians and Jews. In 969 control of the city passed to the Shīʿite Fāṭimid caliphs of Egypt, and in 1010 the emotionally unstable caliph al-Ḥākim ordered the destruction of Christian shrines. In 1071 the Seljuq Turks defeated the Byzantines, displaced the Egyptians as masters of the Holy Land, and cut the pilgrim routes, thus stimulating the Crusades.

The city was recaptured by the Fāṭimids (1098) a year before the hosts of the First Crusade besieged the city. The crusader state took its name from the city, as the Kingdom of Jerusalem. The city regained its position as capital of the kingdom, which (with its northern principalities) stretched from the confines of modern Turkey to the Red Sea. The great Muslim sanctuaries became Christian churches, and in 1149 the Church of the Holy Sepulchre as it exists today was consecrated. Muslims and Jews were barred from living in the city. The kingdom in Jerusalem lasted from 1099 to 1187, when it was overthrown by Saladin, whose Ayyūbid successors ruled from Damascus and Cairo. Jerusalem was again in Christian hands in 1229–39 and 1240–44, when it was sacked by the Khwārezmian Turks. In 1247 the Holy City fell to the Mamlūks, among whom rule of Jerusalem was an object of great desire. The great sanctuaries became Muslim again, and the only Christians who remained were the Greek Orthodox and other Eastern groups. In the 14th century the Franciscans began to represent the Roman Catholic interests. The Jews, who had been barred by the crusaders, returned and from the mid-13th century inhabited their own quarter. The layout of the quarters was fixed in that period. The Mamlūks dotted the Temple Mount and the city with mosques, *madrasah*s, and ornamental tombs.

In 1517 the Ottoman sultan Selim I took the city and inaugurated a Turkish regime that lasted 400 years. The 16th century was a period of great urban development. In addition to the new walls, which still encompass the Old City, and the repaired water supply, new *madrasah*s, *waqf*s, and charity institutions multiplied. But by the end of the century the city began an economic decline that lasted for the next 300 years.

MODERN JERUSALEM

Several factors determined the fate of the city in the 19th century. In 1831 Ibrāhīm Pasha, son of the Egyptian ruler Muḥammad ʿAlī, captured Jerusalem and introduced a series of far-reaching reforms, which were retained when the Turks regained the city (1840). By mid-century all of the great European powers had established consulates in the city, which had a salutary influence on the position of the non-Muslim population. Finally, Jewish immigration, mainly from eastern Europe, changed the city's demographic structure and the relative importance of the Old City and the new quarters outside the walls. Christian and Muslim quarters followed suit. By the mid-19th century nearly half of the city's population was Jewish, and it was expanding beyond the walls.

In 1917 British troops under Sir Edmund Allenby entered Jerusalem after the retreat of the Turks. This opened a new era lasting until 1948, during which Jerusalem again became a capital, now ruled under the British mandate and headed by a high commissioner. About half of its population of some 80,000 was Jewish, with the rest divided between Muslims and Christians. The city developed quickly, expanding its economy and population despite bloody confrontations between Arabs and Jews in 1920, 1929, and 1936–39. With the end of the British mandate in 1947, violence again flared. The UN's attempt to promote an amicable peace between the belligerents failed, and at the end of the Arab-Israeli War of 1948 Jerusalem was divided between the Israelis and the Jordanians. In the Six-Day War of June 1967, the Israelis conquered Jordanian-controlled East Jerusalem and, in the face of international protest, annexed it and large tracts of surrounding territory. Beginning in the late 1970s, Israel engaged in the large-scale development of settlements around Jerusalem, many of which were constructed in occupied territory in the West Bank. Repeated assertions by Israel that Jerusalem was the nation's capital found little international recognition, however, and persistent UN resolutions called for the Jewish state to withdraw from East Jerusalem, which was widely considered to be occupied territory. The issue of Jerusalem remained one of the main points of contention between Israelis and Palestinians, with both parties claiming the city as a capital.

BIBLIOGRAPHY. A good general account of the city is given by TEDDY KOLLEK and MOSHE PEARLMAN in *Jerusalem, Sacred City of Mankind* (1968; also published as *Jerusalem: A History of Forty Centuries*). The best guides to Christian monuments and holy places are EUGENE HOADE, *Guide to the Holy Land,* 11th ed. (1981), and *Jerusalem and Its Environs,* 7th ed. (1966). Noteworthy archaeological works include KATHLEEN M. KENYON, *Jerusalem: Excavating 3000 Years of History* (1967), and *Digging Up Jerusalem* (1974); and the *Encyclopedia of Archaeological Excavations in the Holy Land,* rev. trans. from the Hebrew, vol. 2 (1976), Eng. ed. by MICHAEL AVI-YONAH. For recent excavations and discoveries, see NAHMAN AVIGAD, *Discovering Jerusalem* (1983; originally published in Hebrew, 1980); YIGAEL YADIN (ed.), *Jerusalem Revealed: Archaeology in the Holy City, 1968–1974,* trans. by R. GRAFMAN (1975, reissued 1976); BENJAMIN MAZAR, *The Mountain of the Lord,* trans. from the Hebrew (1975); MEIR BEN-DOV, *In the Shadow of the Temple* (1985); and the *Israel Exploration Journal* (quarterly).

EMIL SCHURER, *The History of the Jewish People in the Age of Jesus Christ (175 B.C.–A.D. 135),* 2 vol. rev. ed. by GEZA VERMES, FERGUS MILLAR, and MATTHEW BLACK (1973–79; originally published in German, 2nd ed., 1886–90), is an early history. GEORGE ADAM SMITH, *Jerusalem: The Topography, Economics and History from the Earliest Times to A.D. 70,* 2 vol. (1907–08, reprinted 1974 in 1 vol.), is a comprehensive survey. The works of the great Dominican scholars HUGHES VINCENT and F.M. ABEL, especially *Jérusalem: recherches de topographie, d'archéologie et d'histoire,* 2 vol. in 4 (1912–26), should also be read. For medieval history, STEVEN RUNCIMAN, *A History of the Crusades,* 3 vol. (1951–54, reprinted 1975), provides essential background; and JOSHUA PRAWER, *The Crusaders' Kingdom: European Colonialism in the Middle Ages* (1972; also published as *The Latin Kingdom of Jerusalem,* 1973), and *Crusader Institutions* (1980), are useful. GUY LE STRANGE (trans.), *Palestine Under the Moslems: A Description of Syria and the Holy Land from A.D. 650 to 1500* (1890, reprinted 1975), is an exhaustive collection of medieval Arabic sources. Also useful is SHLOMO D. GOITEIN, "Jerusalem in the Arab Period (638–1099)," *The Jerusalem Cathedra,* 2:168–196 (1982). T.S.R. BOASE, *Castles and Churches of the Crusading Kingdom* (1967), is a finely illustrated work by an outstanding scholar. AMNON COHEN and BERNARD LEWIS, *Population and Revenue in Towns of Palestine in the Sixteenth Century* (1978); and YEHOSHUA BEN-ARIEH, *Jerusalem in the 19th Century: The Old City* (1984; originally published in Hebrew, 1977), treat the Ottoman period. The beginning of modern scholarship was heralded by three books: EDWARD ROBINSON and ELI SMITH, *Biblical Researches in Palestine, Mount Sinai and Arabia Petraea,* 3 vol. (1841, reprinted 1977), by the founder of scientific biblical geography; JAMES FINN, *Stirring Times; or, Records from Jerusalem Consular Chronicles of 1853 to 1856,* compiled by his widow (1878); and CHARLES W. WILSON *et al., The Recovery of Jerusalem: A Narrative of Exploration and Discovery in the City and the Holy Land* (1871, reprinted 1872), the record of the first underground survey of the ancient city. For the period of the British mandate, ALBERT M. HYAMSON, *Palestine Under the Mandate, 1920–1948* (1950, reprinted 1976), is the standard work. STEWART PERROWNE, *The One Remains* (1954, reissued 1955), is an eyewitness account of the succeeding period. Of particular importance to understanding modern history are MERON BENVENISTI, *Jerusalem, the Torn City* (1976; originally published in Hebrew, 1973); DAVID H.K. AMIRAN, ARIE SHACHAR, and ISRAEL KIMHI (eds.), *Atlas to Jerusalem* (1973), and *Urban Geography of Jerusalem* (1973), a companion volume; and JOEL L. KRAEMER (ed.), *Jerusalem: Problems and Prospects* (1980). General histories are JOHN GRAY, *A History of Jerusalem* (1969), an authoritative survey by a biblical scholar; F.E. PETERS, *Jerusalem: The Holy City in the Eyes of Chroniclers, Visitors, Pilgrims, and Prophets from the Days of Abraham to the Beginnings of Modern Times* (1985, reprinted 1995); K.J. ASALI (ed.), *Jerusalem in History* (1989, reprinted 2000); and A.L. TIBAWI, *Jerusalem: Its Place in Islam and Arab History* (1969). (S.H.P./Jo.Pr.)

Jesus: The Christ and Christology

Jesus Christ, also called Jesus of Galilee or Jesus of Nazareth, was the founder of Christianity, one of the world's largest religions, and the incarnation of God according to most Christians. His teachings and deeds are recorded in the New Testament, which is essentially a theological document and not a biography in the contemporary sense. Details of the historical Jesus, therefore, are difficult to discern, though basic outlines of his career and message can be characterized when considered in the context of 1st-century Judaism and, especially, Jewish eschatology.

This article is divided into the following sections:

Life and teachings

NAME AND TITLE

Ancient Jews usually had only one name, and, when greater specificity was needed, it was customary to add the father's name or the place of origin. Thus, in his lifetime Jesus was called Jesus son of Joseph (Luke 4:22; John 1:45; 6:42), Jesus of Nazareth (Acts 10:38), or Jesus the Nazarene (Mark 1:24; Luke 24:19). After his death, he came to be called Jesus Christ. *Christ* was not originally a name but a title derived from the Greek word *christos*, which translates the Hebrew term *meshiah* (Messiah), meaning "the anointed one." This title indicates that Jesus' followers believed him to be the anointed son of King David, whom some Jews expected to restore the fortunes of Israel. Passages such as Acts of the Apostles 2:36 show that some early Christian writers knew that *the Christ* was properly a title, but in many passages of the New Testament, including those in Paul's letters, the name and the title are combined and used together as Jesus' name: *Jesus Christ* or *Christ Jesus* (Romans 1:1; 3:24). Paul sometimes simply used *Christ* as Jesus' name (*e.g.*, Romans 5:6).

SUMMARY OF JESUS' LIFE

Although born in Bethlehem, according to Matthew and Luke, Jesus was a Galilean from Nazareth, a village near Sepphoris, one of the two major cities of Galilee (Tiberias was the other). He was born to Joseph and Mary shortly before the death of Herod the Great (Matthew 2; Luke 1:5) in 4 BC. According to Matthew and Luke, however, Joseph was only his father legally. They report that Mary was a virgin when Jesus was conceived and that she "was found to be with child from the Holy Spirit" (Matthew 1:18; *cf.* Luke 1:35). Joseph is said to have been a carpenter (Matthew 13:55) and, according to Mark 6:3, Jesus also became a carpenter.

The birth of Jesus

Luke (2:41–52) states that as a child Jesus was precociously learned, but there is no other evidence of his childhood or early life. As a young adult, he went to be baptized by the prophet John the Baptist and shortly thereafter became an itinerant preacher and healer (Mark 1:2–28). In

Francis G. Mayer—Corbis

Detail from "The Baptism of Christ," centre panel of a triptych by Gerard David, c. 1502–07.

his mid-30s, Jesus had a short public career, lasting perhaps less than one year, during which he attracted considerable attention. Some time between AD 29 and 33—possibly AD 30—he went to observe Passover in Jerusalem, where his entrance, according to the Gospels, was triumphant and infused with eschatological significance. While there he was arrested, tried, and executed. His disciples became convinced that he still lived and had appeared to them. They converted others to belief in him, which eventually led to a new religion, Christianity.

JEWISH PALESTINE AT THE TIME OF JESUS

Palestine and the Roman Empire

The political situation. Palestine in Jesus' day was part of the Roman Empire, which controlled its various territories in a number of ways. In the East (eastern Asia Minor, Syria, Palestine, and Egypt), territories were governed either by kings who were "friends and allies" of Rome (often called "client" kings or, more disparagingly, "puppet" kings) or by governors supported by a Roman army. When Jesus was born, all of Jewish Palestine, as well as some of the neighbouring Gentile areas, was ruled by Rome's able "friend and ally" Herod the Great. For Rome, Palestine was important not in itself but because it lay between Syria and Egypt, two of Rome's most valuable possessions. Rome had legions in both countries but not in Palestine. Roman imperial policy required that Palestine be loyal and peaceful, so that it did not undermine Rome's larger interests. This end was achieved for a long time by permitting Herod to remain king of Judaea (37–4 BC) and allowing him a free hand in governing his kingdom, as long as the requirements of stability and loyalty were met.

The Herodians. When Herod died shortly after Jesus' birth, his kingdom was divided into five parts. Most of the Gentile areas were separated from the Jewish areas, which were split between two of Herod's sons, Herod Archelaus, who received Judaea and Idumaea (as well as Samaria, which was non-Jewish), and Herod Antipas, who received Galilee and Peraea. (In the New Testament, Antipas is somewhat confusingly called Herod, as in Luke 23:6–12; apparently the sons of Herod took his name, just as the successors of Julius Caesar were commonly called Caesar.) Both sons were given lesser titles than king: Archelaus was *ethnarch;* Antipas was *tetrarch.* The non-Jewish areas (except Samaria) were assigned to a third son, Philip, to Herod's sister Salome, or to the province of Syria. The emperor Augustus deposed the unsatisfactory Archelaus in AD 6, however, and transformed Judaea, Idumaea, and Samaria from a client kingdom into an "imperial province." Accordingly, he sent a prefect to govern this province. This minor Roman aristocrat (later called a *procurator*) was supported by a small Roman army of approximately 3,000 men. The soldiers, however, came not from Italy but from nearby Gentile cities, especially Caesarea and Sebaste; presumably the officers were from Italy. During Jesus' public career, the Roman prefect was Pontius Pilate (ruled AD 26–36).

Caiaphas and Pontius Pilate

Roman rule. Although nominally in charge of Judaea, Samaria, and Idumaea, the prefect did not govern his area directly; instead, he relied on local leaders. The prefect and his small army lived in the predominantly Gentile city Caesarea, on the Mediterranean coast, about two days' march from Jerusalem. They came to Jerusalem only to ensure peace during the pilgrimage festivals—Passover, Weeks (Shabuoth), and Booths (Sukkoth)—when large crowds and patriotic themes sometimes combined to spark unrest or uprisings. On a day-to-day basis Jerusalem was governed by the high priest. Assisted by a council, he had the difficult task of mediating between the remote Roman prefect and the local populace, which was hostile toward pagans and wanted to be free of foreign interference. His political responsibility was to maintain order and to see that tribute was paid. Caiaphas, the high priest during Jesus' adulthood, held the office from about AD 18 to 36, longer than anyone else during the Roman period, indicating that he was a successful and reliable diplomat. Since he and Pilate were in power together for 10 years, they must have collaborated successfully.

Thus, at the time of Jesus' public career, Galilee was governed by the tetrarch Antipas, who was sovereign within his own domain, provided that he remained loyal to Rome and maintained peace and stability within his borders. Judaea (including Jerusalem) was nominally governed by Pilate, but the actual daily rule of Jerusalem was in the hands of Caiaphas and his council.

Relations between Jewish areas and nearby Gentile areas. Galilee and Judaea, the principal Jewish areas of Palestine, were surrounded by Gentile territories (*i.e.,* Caesarea, Dora, and Ptolemais on the Mediterranean coast; Caesarea Philippi north of Galilee; Hippus and Gadara east of Galilee). There also were two inland Gentile cities on the west side of the Jordan River near Galilee (Scythopolis and Sebaste). The proximity of Gentile and Jewish areas meant that there was some interchange between them, including trade, which explains why Antipas had *telōnēs* (often translated "tax collectors" but more accurately rendered "customs officers") in the villages on his side of the sea of Galilee. There also was some exchange of populations: some Jews lived in Gentile cities, such as Scythopolis, and some Gentiles lived in at least one of the Jewish cities, Tiberias. Jewish merchants and traders could probably speak some Greek, but the primary language of Palestinian Jews was Aramaic (a Semitic language closely related to Hebrew). On the other hand, the Jews resisted paganism and excluded temples for the worship of the gods of Greece and Rome from their cities, along with the Greek educational institutions the *ephebeia* and *gymnasion,* gladiatorial contests, and other buildings or institutions typical of Gentile areas. Because Jewish-Gentile relations in the land that the Jews considered their own were often uneasy, Jewish areas were usually governed separately from Gentile areas. The reign of Herod the Great was the exception to this rule, but even he treated the Jewish and the Gentile parts of his kingdom differently, fostering Greco-Roman culture in Gentile sectors but introducing only very minor aspects of it in Jewish areas.

Roman Jewish policy

In the 1st century, Rome showed no interest in making the Jews in Palestine and other parts of the empire conform to common Greco-Roman culture. A series of decrees by Julius Caesar, Augustus, the Roman Senate, and various city councils permitted Jews to keep their own customs, even when they were antithetical to Greco-Roman culture. For example, in respect for Jewish observance of the Sabbath, Rome exempted Jews from conscription in Rome's armies. Neither did Rome colonize Jewish Palestine. Augustus established colonies elsewhere (in southern France, Spain, North Africa, and Asia Minor), but prior to the First Jewish Revolt (AD 66–74) Rome established no colonies in Jewish Palestine. Few individual Gentiles from abroad would have been attracted to live in Jewish cities, where they would have been cut off from their customary worship and cultural activities. The Gentiles who lived in Tiberias and other Jewish cities were probably natives of nearby Gentile cities, and many were Syrians, who could probably speak both Aramaic and Greek.

Economic conditions. Most people in the ancient world produced food, clothing, or both and could afford few luxuries. Most Palestinian Jewish farmers and herdsmen, however, earned enough to support their families, pay their taxes, offer sacrifices during one or more annual festivals, and let their land lie fallow in the sabbatical years, when cultivation was prohibited. Galilee in particular was relatively prosperous, since the land and climate permitted abundant harvests and supported many sheep. Although it is doubtful that Galilee was as affluent in the 1st century as it was during the late Roman and Byzantine periods, archaeological remains from the 3rd, 4th, and 5th centuries nevertheless confirm the plausibility of 1st-century references to the region's prosperity. There were, of course, landless people, but the Herodian dynasty was careful to organize large public works projects that employed thousands of men. Desperate poverty was present, too, but never reached a socially dangerous level. At the other end of the economic spectrum, few if any Palestinian Jews had the vast fortunes that successful merchants in port cities could accumulate; however, there were Jewish aristocrats with large estates and grand houses, and the merchants who served the Temple (supplying, for example, incense and fabric) could become very prosperous. The gap between rich and poor in Palestine was obvious and distressing to the poor, but compared with that of the rest of the world it was not especially wide.

The Jewish religion in the 1st century. Judaism, as the Jewish religion came to be known in the 1st century AD, was based on ancient Israelite religion, shorn of many of its Canaanite characteristics but with the addition of important features from Babylonia and Persia. The Jews differed from other peoples in the ancient world because they believed that there was only one God. Like other peoples, they worshiped their God with animal sacrifices offered at a temple; unlike others, they had only one temple, which was in Jerusalem. The sanctuary of the Jewish temple had two rooms, as did many of the other temples in the ancient world, but the second room of the Jewish temple was empty. There was no idol representing the God of Israel. The Jews also believed that they had been specially chosen by the one God of the universe to serve him and obey his laws. Although set apart from other peoples, they believed God called on them to be a "light to the Gentiles" and lead them to accept the God of Israel as the only God.

The Torah. An important part of Jewish Scripture was the Torah, or Pentateuch, comprising five books (Genesis, Exodus, Leviticus, Numbers, and Deuteronomy) that were believed to have been given to Moses by God. For Jews and their spiritual descendants, these books contain God's law, which covers many aspects of ordinary life: it requires that males be circumcised, regulates diet, mandates days of rest for humans and animals alike, requires pilgrimage and sacrifice, stipulates recompense and atonement following transgression, and specifies impurities and required purification before entry to the Temple. Moreover, it provides both rules and principles for the treatment of other people: for example, calling for the use of honest weights and measures in trade and for "love" (that is, upright treatment) of both fellow Jews and foreigners (Leviticus 19). The laws governing worship were similar to the religious laws of other peoples in the ancient world. Judaism was different, however, because it regulated not only worship but also daily life and made every aspect of life a matter of divine concern.

Religious parties. Since both faith and practice were based firmly on the five books of Moses modified slightly over time, they were shared by Jews all over the world, from Mesopotamia to Italy and beyond. The common features of Jewish faith and practice are reflected in the decrees from various parts of the ancient world that allowed Jews to preserve their own traditions, including monotheism, rest and assembly on the Sabbath, support of the Temple, and dietary laws. There were, naturally, variations on each main theme. In Jewish Palestine, for example, there were three small but important religious parties that differed from each other in several ways: the Pharisees (numbering about 6,000 at the time of Herod), Essenes (about 4,000), and Sadducees ("a few men," according to Flavius Josephus, in *The Antiquities of the Jews* 18.17). A largely lay group that had the reputation of being the most precise interpreters of the law, the Pharisees believed in the resurrection of the dead. They also relied on the nonbiblical "traditions of the fathers," some of which made the law stricter while others relaxed it. The Essenes were a more radical sect, with extremely strict rules. One branch of the group lived at Qumrān on the shores of the Dead Sea and produced the Dead Sea Scrolls. At some point in their history the Essenes were probably a priestly sect (the Zadokite priests are major figures in some of the documents from Qumrān); however, the composition of their membership at the time of Jesus is unclear. Many aristocratic priests, as well as some prominent laymen, were Sadducees. They rejected the Pharisaic "traditions of the fathers" and maintained some old-fashioned theological opinions; most famously, they denied resurrection, which had recently entered Jewish thought from Persia and which was accepted by most Jews in the 1st century.

Pharisees, Sadducees, and Essenes

Most Jews based their faith and practice on the five books of Moses (slightly modified by the passage of time) and rejected the extreme positions of the three parties. The Pharisees were respected for their piety and learning, and they may have exercised substantial influence on belief and practice. The Essenes were a fringe group, and those who lived at Qumrān had dropped out of mainstream Judaism. Their interpretation of the Bible led them to reject the priests and the Temple as they existed in Jerusalem, and they looked forward to the time when they could seize control of the Holy City. To the degree that any of these parties had power, however, it belonged to the Sadducees. More precisely, the aristocratic priests and a few prominent laymen had power and authority in Jerusalem; of the aristocrats who belonged to one of the parties, most were Sadducees. According to the Acts of the Apostles (5:17), those who were around the high priest Caiaphas were Sadducees, which recalls the evidence of the Jewish priestly aristocrat, historian, and Pharisee Josephus.

While the majority of Jews did not belong to a party, the study of these parties reveals the variety within the framework of Judaism. Another indicator of this variety is the diversity of Jewish leaders. Among them were charismatic healers and miracle workers, such as Honi the Circle Drawer; hermitlike sages, such as Bannus; eschatological prophets, such as John the Baptist; would-be messianic prophets, such as Theudas; and apocalyptic visionaries, represented by the pseudepigraphal First Book of Enoch.

Eschatology. Most Jews had some form of future hope; they expected God to intervene in history and to restore Israel to a state of peace, freedom, and prosperity. Not all Jews expected God to send a son of David as Messiah to overthrow the Romans, though some did. The Qumrān sect believed that there would be a great war against Rome, that the sect would emerge victorious, and that the main blows would be struck by the angel Michael and finally by God himself. Notably, a Messiah plays no role in this war of liberation. Some Jews were ready at any moment to take up arms against Rome, thinking that if they started the fight God would intervene on their side. Others were quietists, hoping for divine deliverance but unwilling to fight. Whatever their expectations, few Palestinian Jews were completely satisfied with the governments of Antipas, Pilate, and Caiaphas. As God's chosen people, the Jews felt they should be free both of foreign domination and of ambitious worldly leaders.

In the final analysis, variety and commonality are equally important to the understanding of Palestinian Judaism in Jesus' day. Jews agreed on many aspects of their religion, and they agreed that they did not want to surrender their covenant with God to accept the lure of pagan culture; but, when it came to details, they could disagree with one another violently. Since God cared about every aspect of life, competing groups and leaders often saw themselves as representing the side of God against his adversaries.

SOURCES FOR THE LIFE OF JESUS

Apocryphal gospels. The only substantial sources for the life and message of Jesus are the Gospels of the New Testament, the earliest of which was Mark (written AD 60–80), followed by Matthew, Luke, and John (AD 75–90). Some evidence can be found in the letters of Paul, which were written beginning in AD 50 and are the earliest surviving Christian texts. There are other sources that may have information. Notably, an ossuary (a box to hold the bones of the dead) was discovered from AD 63 with the inscription "James, son of Joseph, brother of Jesus." If authentic, as some scholars believe, and if the "Jesus" referred to is indeed Jesus Christ, this would constitute the earliest artifact to mention him. The ossuary was the subject of much commentary when it was made public in 2002.

Noncanonical sources, especially the apocryphal gospels, contain many sayings attributed to Jesus, as well as stories about him that are occasionally held to be "authentic." Among these apocrypha, the Gospel of Thomas (written in the mid-2nd century AD) has attracted much attention. A "sayings" gospel (114 sayings attributed to Jesus, without narrative), it is grounded in Gnosticism, a philosophical and religious movement of the 2nd century AD that stressed the redemptive power of esoteric knowledge acquired by divine revelation. For Thomas, salvation consists of self-knowledge, and baptism results in restoration to the primordial state—man and woman in one person, like Adam before the creation of Eve (saying 23). Spiritual reversion to this

The Gospel of Thomas

state meant that nakedness need not result in shame; one passage (saying 37) allows us to suspect that the early Christian followers of the Gospel of Thomas took off their garments and trampled on them as part of their baptismal initiation. There are a few connections between this world-view and that of Paul and the Gospel According to John, but the overall theology of the Gospel of Thomas is so far removed from the teaching of Jesus as found in the Gospels of Matthew, Mark, and Luke—in which Jewish eschatology is central—that it is not considered a major source for the study of Jesus. It is, of course, possible or even likely that individual sayings in Thomas or other apocryphal gospels originated with Jesus, but it is unlikely that noncanonical sources can contribute much to the portrait of the historical Jesus. As in the case of the Gospel of Thomas, the traditions found in other apocryphal gospels are often completely unlike the evidence of the canonical gospels and are embedded in documents that are generally believed to be unreliable.

Roman and Jewish sources. There are a few references to Jesus in 1st-century Roman and Jewish sources. Documents indicate that within a few years of Jesus' death, Romans were aware that someone named Chrestus (a slight misspelling of Christus) had been responsible for disturbances in the Jewish community in Rome (Suetonius, *The Life of the Deified Claudius* 25.4). Twenty years later, according to Tacitus, Christians in Rome were prominent enough to be persecuted by Nero, and it was known that they were devoted to Christus, whom Pilate had executed (*Annals* 15.44). This knowledge of Jesus, however, was dependent on familiarity with early Christianity and does not provide independent evidence about Jesus. Josephus wrote a paragraph about Jesus (*The Antiquities of the Jews* 18.63ff.), as he did about Theudas, the Egyptian, and other charismatic leaders (*History of the Jewish War* 2.258–263; *The Antiquities of the Jews* 20.97–99, 167–172), but it has been heavily revised by Christian scribes, and Josephus' original remarks cannot be discerned.

The Pauline Letters. The letters of Paul contain reliable but meagre evidence. Their main theme, that Jesus was crucified and raised from the dead, is especially prominent in 1 Corinthians 15, where Paul evokes an early tradition about Jesus' death and subsequent appearances to his followers. The Crucifixion and Resurrection were accepted by all first-generation Christians. Paul also quotes a few of Jesus' sayings: the prohibition of divorce and remarriage (1 Corinthians 7:10–11), the words over the bread and cup at Jesus' Last Supper (1 Corinthians 11:23–25), and a prediction of the imminent arrival of the Saviour from heaven (1 Thessalonians 4:15–17).

The Gospels. Fuller information about Jesus is found in the Gospels of the New Testament, though these are not of equal value in reconstructing his life and teaching. The Gospels of Matthew, Mark, and Luke agree so closely with one another that they can be studied together in parallel columns in a work called a synopsis and are hence called the Synoptic Gospels. John, however, is so different that it cannot be reconciled with the Synoptics except in very general ways (*e.g.,* Jesus lived in Palestine, taught, healed, was crucified and raised). In the Synoptics, Jesus' public career appears to have lasted less than one year, since only one Passover is mentioned; in John, three Passovers occur, implying a ministry of more than two years. In all four Gospels, Jesus performs miracles, especially healings, but, while exorcisms are prevalent in the Synoptics, there are none in John. The greatest differences, though, appear in the methods and content of Jesus' teaching. In the Synoptic Gospels, he speaks about the kingdom of God in short aphorisms and parables, making use of similes and figures of speech, many drawn from agricultural and village life. He seldom refers to himself, and, when asked for a "sign" to prove his authority, he refuses (Mark 8:11–12). In John, on the other hand, Jesus employs long metaphorical discourses, in which he himself is the main subject. His miracles are described as "signs" that support the authenticity of his claims.

John and the Synoptic Gospels

The Synoptic Gospels. Scholars have unanimously chosen the Synoptic Gospels' version of Jesus' teaching. The verdict on the miracles is the same, though less firmly held: in all probability Jesus was known as an exorcist, which resulted in the charge that he cast out demons by the prince of demons (Mark 3:22–27). The choice between the narrative outline of the Synoptics and that of John is less clear. Besides presenting a longer ministry than do the other Gospels, John also describes several trips to Jerusalem. Only one is mentioned in the Synoptics. Both outlines are plausible, but a ministry of more than two years leaves more questions unanswered than does one of a few months. It is generally accepted that Jesus and his disciples were itinerant; that they traveled around Galilee and its immediate environs; and that Jesus taught and healed in various towns and villages, as well as in the countryside and on the shore of the Sea of Galilee. But where did they spend their winters? Who supported them? None of the Gospels explains how they lived (though Luke 8:1–3 alludes to some female supporters), but the omission is even more glaring in John, where the longer ministry presumes the need for winter quarters, though none are mentioned. This and other considerations are not decisive, but the brief career of the Synoptic Gospels is slightly to be preferred.

The Synoptic Gospels, then, are the primary sources for knowledge of the historical Jesus. They are not, however, the equivalent of an academic biography of a recent historical figure; instead, the Synoptic Gospels are theological documents that provide information the authors regarded as necessary for the religious development of the Christian communities in which they worked. The details of Jesus' daily life are almost entirely lacking, as are such important features as his education, travel, and other developmental experiences. The characters on the whole are "flat": emotions, motives, and personalities are seldom mentioned. There are, nevertheless, a few exceptions that show how little is actually known. Peter wavers (Matthew 14:28–31; Mark 14:66–72); James and John ask for preferential treatment in the coming kingdom (Mark 10:35–40); and Pilate anguishes over the decision to execute Jesus (Matthew 27:15–23; Luke 23:2–25). On the other hand, the Pharisees and scribes periodically challenge Jesus and then disappear, with little indication of what, from their point of view, they hoped to accomplish. Even Jesus is a rather flat character in the Gospels. He is sometimes angry and sometimes compassionate (Mark 3:5; 6:34, respectively), but one can say little more. This is a frustrating aspect of the Gospels. The situation is different with regard to Paul, whose letters are extant and self-revelatory. The force of his personality is in the letters, but the force of Jesus' personality must be found somewhere behind the Gospels.

The Gospels comprise brief, self-contained passages, or *pericopēs* (from the Greek word meaning "cut around"), relating to Jesus. Further study reveals that the authors of the Synoptic Gospels moved these pericopes around, altering their contexts to suit their own editorial policies, for example by arranging the pericopes according to subject matter. In chapters 8 and 9, Matthew collects 10 healing pericopes, with a few other passages interspersed. Mark and Luke contain most of these passages, but their arrangements are different. Table 1 delineates the Synoptics' accounts of 10 of Jesus' healings. Matthew put all of these healings in one place; Mark and Luke scattered them, but in different ways. Since the authors of the Gospels rearranged the material to suit their own needs, it must be assumed that earlier Christian teachers had also organized stories about Jesus didactically. This means that the sequence of events in Jesus' ministry is unknown.

Arrangement of the Gospels

Moreover, the Evangelists and other early Christian teachers also shaped the material about Jesus. During the course of transmission, the factual narrative elements that surrounded each saying or event were stripped away, leaving only a central unit, which was applied to various situations by the addition of new introductions and conclusions. For example, both Matthew and Luke relate the Parable of the Lost Sheep. In Matthew 18:12–14, the parable is told to the disciples, and the meaning is that they, like the shepherd, should go in search of the lost. In Luke 15:4–7, the same story is directed at the Pharisees, this time to instruct them not to grumble because Jesus has attracted repentant sinners. Both applications of the parable were

Table 1: The placement of 10 of Jesus' healings in the Synoptic Gospels

	title of pericope	Matthew	Mark	Luke
1.	The leper	8:1–4	1:40–45	5:12–16
2.	Centurion's servant	8:5–10	—	7:1–10
3.	Peter's mother-in-law	8:14–15	1:29–31	4:38–39
4.	Healing summary Discipleship Calming storm	8:16–17 8:18–22 8:23–27	1:32–34 — 4:35–41	4:40–41 9:57–60 8:22–25
5.	Gadarene demoniacs	8:28–34	5:1–20	8:26–39
6.	The Paralytic Call of Levi Fasting	9:1–8 9:9–13 9:14–17	2:1–12 2:13–17 2:18–22	5:17–26 5:27–32 5:33–39
7,8.	Jairus' daughter; woman with hemorrhage	9:18–26	5:21–43	8:40–56
9.	Blind men	9:27–31	—	—
10.	Demoniac	9:32–34	—	—

useful homiletically, and therefore were preserved. The context in which Jesus originally used the parable, however, is unknown. Another example is the saying "love your enemies" (Matthew 5:44). Homiletically, it may be applied to numerous circumstances, which makes it very useful for sermons and teaching. Historically, however, it is not known to whom Jesus referred when he spoke these words. The lack of firm knowledge of original context makes the precise interpretation of individual passages difficult.

Authenticity. Further, not all the sayings and deeds in the Synoptic Gospels are reports of things that Jesus actually said and did. Believing that Jesus still lived in heaven, the early Christians spoke to him in prayer and sometimes he answered (2 Corinthians 12:8–9; *cf.* 1 Corinthians 2:13). These early Christians did not distinguish between "the historical Jesus" and "the heavenly Lord" as firmly as most modern people do, and some sayings heard in prayer almost certainly ended up in the Gospels as sayings uttered by Jesus during his lifetime.

Since both the original context of Jesus' sayings and deeds and those passages in the Gospels that go back to the historical Jesus are unknown, there are substantial difficulties in attempting to reconstruct the Jesus of history. Of these two difficulties, the lack of immediate context is the more serious. It must be admitted that, on many points, precision and nuance in describing the teaching and ministry of Jesus cannot be achieved.

Tests of authenticity

There are, however, tests of authenticity that make it possible to acquire good general information about Jesus' teachings. One of the most important of these is "multiple attestation": a passage that appears in two or more independent sources is likely to be authentic. A prime example is the prohibition of divorce, which appears in the letters of Paul and in two different forms in the Synoptic Gospels. The short form, which is focused on remarriage after divorce, is found in Matthew 5:31–32 and Luke 16:18. The long form, which is more absolute in prohibiting divorce, appears in Matthew 19:1–12 and Mark 10:1–12. Paul's version (1 Corinthians 7:10–11) agrees most closely with the short form. Because of this excellent attestation, it is almost indisputable that Jesus opposed divorce and especially remarriage after divorce, though study of the five passages does not reveal precisely what he said.

A second test is "against the grain of the Gospels": a passage that seems to be contrary to one of the main themes or views expressed in one or more Gospels is likely to be authentic because the early Christians were not likely to have created material with which they disagreed. Matthew's depiction of John the Baptist is a good example. The author apparently found it to be embarrassing that Jesus received John's baptism of repentance (why would Jesus have needed it?). Thus, he has John protest against the baptism and claim that Jesus should instead baptize him (Matthew 3:13–17; this objection is not in Mark or Luke). These verses in Matthew assume that John recognized Jesus as being greater than he, but Matthew later shows John, in prison, sending a message to ask Jesus whether he was "the one who is to come" (Matthew 11:2–6). These passages make it virtually certain that John baptized Jesus and highly probable that John asked Jesus who he was. John's protest against baptizing Jesus appears to be Matthew's creation. In keeping these passages while, in effect, arguing against them, Matthew validates the authenticity of the tradition that John baptized Jesus and later enquired about his true identity.

These are only a few examples of tests that may confirm the authenticity of passages in the Gospels. In many cases, however, the criteria do not apply: many passages neither meet nor fail the tests. Grouping passages into categories—probable, improbable, possible but unconfirmed—is a useful exercise but does not go very far toward determining a realistic portrayal of Jesus as a historical figure. More is needed than just the minute study of the Gospels, though that is an essential task.

THE CONTEXT OF JESUS' CAREER

Good historical information about Jesus can be acquired by establishing the overall context of his public ministry. As noted earlier, he began his career by being baptized by John, an eschatological prophet, and an understanding of eschatology is pivotal to interpreting Jesus' world. Although eschatology is the doctrine of last things, the Jews who anticipated future redemption did not expect the end of the world. Instead, they thought that God would intervene in human history and make the world perfect: that is, the Jews would live in the Holy Land free of foreign domination and in peace and prosperity. Many Jews, including John, expected final judgment to precede this golden age, and he taught that people should repent in view of its imminence (Matthew 3:1–12; Luke 3:3–9). Since Jesus accepted John's baptism, he must have agreed with this message, at least in part. After Jesus' death and Resurrection, his followers believed that he would soon return to bring in the kingdom of God. The clearest expression of this belief is offered by Paul, whose earliest letter indicates that the Lord will return before most of the people then alive die (1 Thessalonians 4:13–18). If Jesus began his career by being baptized by an eschatological prophet and if after his Crucifixion his followers expected him to return to save them (1 Thessalonians 1:9–10; 1 Corinthians 15:20–28), it is highly probable that he himself shared the basic views of Jewish eschatology.

Jesus and eschatology

Many aspects of Jesus' career support the view that he expected divine intervention. One of the most common beliefs of Jewish eschatology was that God would restore the Twelve Tribes of Israel, including the Ten Lost Tribes. That Jesus shared this view is indicated by his call of 12 disciples, who apparently represented the 12 tribes (Matthew 19:28). Moreover, he proclaimed the arrival of the kingdom of God; he predicted the destruction of the Temple (Mark 13:2) and possibly its rebuilding "without hands" (Mark 14:58); he entered Jerusalem on a donkey, symbolizing his kingship (Mark 11:4–8; Matthew 21:1–11; see Zechariah 9:9 for the symbol); and he had a final meal with his disciples in which he said that he would "drink no more of the fruit of the vine until that day when I drink it in the new kingdom of God" (Mark 14:25). It is no surprise that after his death his disciples formed a small community that expected Jesus to return and inaugurate a kingdom in which the world would be transformed.

In this light, Jesus can be seen as an eschatological prophet, grouped historically in the same general category as John the Baptist and a few other 1st-century Jewish prophets, such as Theudas. Like John, Jesus believed in the coming judgment, but he stressed inclusion more than condemnation and welcomed "customs officers and sinners" in the coming kingdom of God (Matthew 11:18–19; 21:31–32). Moreover, his teaching was rich and multifaceted and was not limited to eschatological expectation.

MAIN ASPECTS OF JESUS' TEACHING

The kingdom of God. While the Gospels agree that Jesus proclaimed the eschatological kingdom of God, they offer different versions of his view of that kingdom. One is that the kingdom of God exists in heaven and that individuals may enter it upon death (Mark 9:47). Since God's power is in some respects omnipresent, Jesus may have seen "the kingdom," in the sense of God's presence, as being especially evident in his own words and deeds. The parable that the kingdom is like yeast that gradually leavens the entire loaf (Matthew 13:33) indicates that Jesus may have understood the kingdom of God to be beginning in the present. These other ways of viewing the kingdom do not, however, dominate the teaching of Jesus in the Synoptic Gospels. Statements about the heavenly kingdom, or the kingdom as partially present on earth, do not negate the eschatological nature of Jesus' message. The essence of his teaching is that the kingdom would come to earth in its full power and glory, at which time God's will would be done "on earth as it is in heaven" (Matthew 6:10). Jesus died before heaven came down to earth, and this, coupled with the Resurrection appearances, led his followers to expect him to return in the near future, ushering in the kingdom and ruling in God's stead.

The Son of Man

Jesus himself apparently anticipated the arrival of a heavenly figure whom he called "the Son of Man," who would come on clouds of glory and gather the elect. The Hebrew Bible laid the foundation for this teaching in two ways. First, several prophets expected "the day of the Lord," when the wicked would be punished or destroyed and the good would be spared, though the emphasis was on punishment (Amos 5:12–20; Zephaniah 1; Joel 1:15; 2:1; Obadiah verse 15). Second, Daniel 7 describes various kingdoms that are represented by four fantastic beasts, all of which are destroyed. Then, according to Daniel, the Son of Man, representing the people of Israel, ascends to God and receives "dominion and glory and kingship" (Daniel 7:14), after which Israel is to reign supreme (7:27). These passages seem to have led Jesus to depict the arrival of the Son of Man from heaven as initiating the coming judgment and the redemption of Israel. The theme appears in numerous passages in the Synoptic Gospels (see Tables 2, 3, 4, and 5).

Table 2: Cosmic distress, Son of Man, angels, gather elect
Matthew 24:29–31
Immediately after the suffering of those days the sun will be darkened, and the moon will not give its light; the stars will fall from heaven, and the powers of heaven will be shaken. [30]Then the sign of the Son of man will appear in heaven, and then all the tribes of the earth will mourn, and they will see "the Son of Man coming on the clouds of heaven" with power and great glory. [31]And he will send out his angels with a loud trumpet call, and they will gather his elect from the four winds, from one end of heaven to the other.
Mark 13:24–27
But in those days, after that suffering, the sun will be darkened, and the moon will not give its light, [25]and the stars will be falling from heaven, and the powers in the heavens will be shaken. [26]Then they will see "the Son of Man coming in clouds" with power and glory. [27]Then he will send out the angels, and gather his elect from the four winds, from the ends of the earth to the ends of heaven.
Luke 21:25–28
There will be signs in the sun, the moon, and the stars, and on the earth distress among nations confused by the roaring of the sea and the waves. [26]People will faint from fear and foreboding of what is coming upon the world, for the powers of the heavens will be shaken. [27]Then they will see "the Son of Man coming in a cloud" with power and great glory. [28]Now when these things begin to take place, stand up and raise your heads, because your redemption is drawing near.

Paul's depiction of the coming kingdom also merits consideration (italics indicate the closest agreements with the passages in the Gospels):

For this we declare to you by the word of the Lord, that *we who are alive,* who are left until the appearance of the Lord, will not precede those who have fallen asleep. For the Lord himself *will come down from heaven with a command, with the voice of an archangel, and with a trumpet of God;* and the dead in Christ will rise first. *Then we who are alive, who are left, will be caught up in the clouds* together with them to meet the Lord in the air (1 Thessalonians 4:15–17).

Table 3: Son of Man, glory of Father, angels, some will not taste death
Matthew 16:27–28
For the Son of Man is to come with his angels in the glory of his Father, and then he will repay everyone for what has been done. [28]Truly I tell you, there are some standing here who will not taste death before they see the Son of Man coming in his kingdom.
Mark 8:38–9:1
Those who are ashamed of me and of my words in this adulterous and sinful generation, of them the Son of Man will also be ashamed when he comes in the glory of his Father with the holy angels. [9:1]And he said to them, Truly I tell you, there are some standing here who will not taste death until they see that the kingdom of God has come with power.
Luke 9:26–27
Those who are ashamed of me and of my words, of them the Son of Man will be ashamed when he comes in his glory and the glory of the Father and of the holy angels. [27]But truly I tell you, there are some standing here who will not taste death before they see the kingdom of God.

Table 4: Son of Man comes like lightning or the great flood
Matthew 24:27, 37–39
For as the lightning comes from the east and flashes as far as the west, so will be the coming of the Son of Man....[37]For as the days of Noah were, so will be the coming of the Son of Man. [38]For as in those days before the flood they were eating and drinking, marrying and giving in marriage, until the day Noah entered the ark, [39]and they knew nothing until the flood came and swept them all away, so too will be the coming of the Son of Man.
Luke 17:24, 26, 27, 30
For as the lightning flashes and lights up the sky from one side to the other, so will the Son of Man be in his day....[26]Just as it was in the days of Noah, so too it will be in the days of the Son of Man. [27]They were eating and drinking, and marrying and being given in marriage, until the day Noah entered the ark, and the flood came and destroyed all of them....[30][I]t will be like that on the day that the Son of Man is revealed.

Table 5: Preparedness
Matthew 24:44
Therefore you also must be ready, for the Son of Man is coming at an unexpected hour.
Luke 12:40
You also must be ready, for the Son of Man is coming at an unexpected hour.

Paul changed "the Son of Man" to "the Lord." It is not known whether Jesus intended to refer to himself or to another figure when he used the term *Son of Man* in this context (he did refer to himself as a Son of Man in the sense of "a human being," as in Matthew 8:20). By Paul's time, however, Christians made no such distinction and interpreted the heavenly Son of Man as the risen Jesus.

Jesus' belief that the Son of Man would soon arrive to usher in the kingdom is confirmed as authentic by multiple attestation. It is also "against the grain" of the Gospel According to Luke, since the author tended to downplay eschatology (*e.g.,* Luke 17:21 and Acts, written by the same author). Moreover, Paul, whose letters are earlier than the gospels, thought that most people then living would still be alive at the time of Jesus' return, whereas the Synoptic Gospels state that "some standing here will not taste death." The change from "most" to "some" probably demonstrates that the expectation was beginning to fade when the Gospels were written.

Inclusion in the kingdom. Several passages indicate that following Jesus was highly desirable for those who wished to be included in the coming kingdom. Jesus called on some people to give up everything in order to follow him (Mark 1:16–20; 10:17–31) and promised that their reward would be great in heaven. It cannot be said, however, that Jesus viewed personal loyalty as a prerequisite for inclusion in the kingdom. Often he simply urged all to fix their attention on the kingdom, not on material possessions (Matthew 6:19–21; 6:25–34; Luke 12:13–21). The majority of his teaching in the Synoptic Gospels is about God and the value of returning to him. Thus, in one parable the "prodigal son" returns to "the father," presumably representing God (Luke 15:11–32).

The kingdom and the poor in spirit

Perhaps faith in God and treating other people with love (Matthew 25:34–40) would suffice for entry into the kingdom. This seems to be indicated by the study of children, the childlike, the poor, the meek, the lowly, and the sinners, whom Jesus especially called and favoured. "Let the little children come to me; do not stop them; for it is to such as these that the kingdom of God belongs" (Mark 10:14). In the coming kingdom, moreover, the last would be first (Mark 10:31); those who held the chief positions in the present world would be demoted (Luke 14:7–11); those who gave up everything and followed Jesus would receive "a hundred-fold" (Mark 10:30); and sinners, exemplified by the customs officers and prostitutes, also would be included in the kingdom (Matthew 21:31). The Beatitudes, drawn from the Sermon on the Mount, particularly stress Jesus' concern for the poor and the meek who will be blessed (Luke 6:20; Matthew 5:3–5). This emphasis probably rests in part on his sympathy for those of his own socioeconomic class or below it. Significantly, Jesus and his disciples were not themselves from the very bottom of society. His father worked with his hands, but he was not destitute, and some of Jesus' disciples were from families who owned fishing boats and houses (Mark 1:19, 29). They were not rich, but they also were not day labourers, beggars, or homeless, all of whom were the focus of Jesus' sympathy.

His message had a social dimension in two respects. He thought that in the kingdom there would be social relationships, not a collection of disembodied spirits floating on the clouds. He also believed that the disadvantaged of the present world would be in some sense or other advantaged in the new age (Matthew 5:3–11; Luke 6:20–23). It is possible that the promise of houses and lands in Matthew 19:29 and Mark 10:29–30 is metaphorical, but it is also possible that Jesus envisaged a future society in which property would still count, though it would be redistributed.

Jesus' appeal to sinners, according to Luke 5:32, meant that he called them to repent, but neither Matthew 9:13 nor Mark 2:17 mentions the word *repentance.* Most likely, Jesus' message was more radical than a simple call for repentance, a proposition with which everyone would have agreed. He wanted sinners to accept him and his message, and he promised inclusion in the kingdom if they did so. This acceptance doubtless included moral reformation, but Jesus probably did not mean that they had to conform precisely to the standards of righteous Jewish society, which demanded repayment of money or goods obtained dishonestly, the addition of one-fifth as a fine, and the presentation of a guilt-offering in the Temple (Leviticus 6:1–7). Instead, Jesus called people to follow him and to be like his disciples. He evidently expected more people to be like him (accepting sinners, loving even enemies) than to join the small band that followed him. Although Jesus specifically called several followers, he seems not to have viewed personal faith in and commitment to him as absolute necessities (though faith in him became the standard requirement of early Christianity).

Jesus' use of messianic titles

Whether he made this particular requirement or not, Jesus certainly attached great importance to his own mission and person. The Christian preoccupation with titles (did he think that he was the Messiah, Son of God, heavenly Son of Man, son of David, or king?) obscures the issue. Jesus sometimes called himself the Son of Man, though perhaps not meaning the heavenly Son of Man, and according to two passages he indirectly accepted the epithets Messiah (or Christ) and Son of God (Matthew 16:16; Mark 14:61–62). In both cases, however, the parallel passages (Mark 8:29; Luke 9:20; 22:67–70; Matthew 26:63–64) are less strongly affirmative. In any case, Jesus apparently did not make an issue of titles. He called people to follow him and to devote themselves entirely to God, not to accord him a particular appellation. If he was preoccupied with titles, the evidence is so scant that it cannot be known what those titles meant to him or others. If, however, this uncertain evidence is overlooked, a clearer picture of his self-conception emerges: Jesus thought that he was God's last emissary, that he and his disciples would rule in the coming kingdom, and that people who accepted his message would be included in it. He may also have believed that inclusion in the kingdom would be granted to those who loved their neighbours and were meek and lowly of heart.

The relation of Jesus' teaching to the Jewish law. Jewish law is the focus of many passages in the Gospels. According to one set, especially prominent in the Sermon on the Mount (Matthew 5–7), Jesus admonished his followers to observe the law unwaveringly (Matthew 5:17–48). According to another set, he did not adhere strictly to the law himself and even transgressed current opinions about some aspects of it, especially the Sabbath (*e.g.,* Mark 3:1–5). It is conceivable that both were true, that he was extremely strict about marriage and divorce (Matthew 5:31–32; Mark 10:2–12) but less stringent about the Sabbath. The study of Jesus and the law is, like any other study of law, highly technical. In general, the legal disputes in the Gospels fall within the parameters of those of 1st-century Judaism. Some opposed minor healing on the Sabbath (such as Jesus is depicted as performing), but others permitted it. Similarly, the Sadducees regarded the Pharisees' observance of the Sabbath as too lax. There also were many disagreements in 1st-century Judaism about purity. While some Jews washed their hands before eating (Mark 7:5), others did not; however, this conflict was not nearly as serious as that between the Shammaites and the Hillelites (the two main parties within Pharisaism) over menstrual purity. It is noteworthy that Jesus did not oppose the purity laws. On the contrary, according to Mark 1:40–44, he accepted the Mosaic laws on the purification of lepers (Leviticus 14).

The Sabbath, diet, and autonomy

In one statement in the Gospels, however, Jesus apparently opposed Jewish law as universally understood. Jews agreed not to eat carnivores, rodents, insects, and weasels, as well as pork and shellfish (Leviticus 11; Deuteronomy 14), and the last two prohibitions set them apart from other people. According to Mark 7:19, Jesus "declared all foods clean." If he did so, Jesus directly opposed the law of God as given to Moses. This seems to be only Mark's inference, however, and is not in the parallel passage in Matthew 15. More importantly, Peter seems to have first learned of this after Jesus' death, by means of a heavenly revelation (Acts 10:9–16). Perhaps Jesus did not, then, directly oppose any aspect of the sacred law.

He probably did, however, have legal disputes in which he defended himself by quoting scriptural precedent, which implies that he did not set himself against the law (Mark 2:23–28). His willingness to make his own decisions regarding the law was probably viewed with suspicion. Ordinarily, legal debates were between competing camps or schools, and individuals who decided how to observe laws were deemed troublemakers. That is, Jesus was *autonomous;* he interpreted the law according to his own rules and decided how to defend himself when criticized. He was by no means the only person in ancient Judaism who struck out on his own, acting in accord with his own perception of God's will, and so he was not uniquely troubling in this respect, but such behaviour might nevertheless be suspicious.

Ethics. Along with his teachings on the kingdom and the law, Jesus advocated ethical purity. He demanded complete devotion to God, putting it ahead of devotion to self and even to family (Mark 3:31–35; Matthew 10:35–37), and taught that people should give up everything in order to obtain what was most precious (Matthew 13:44–46). According to Matthew 5:21–26 and 5:27–30, Jesus also held that observance of the law should be not only external but internal: hatred and lust, as well as murder and adultery, are wrong. The Jesus of Matthew in particular is a moral perfectionist (5:17–48). This fits quite well with the proclamation of the eschatological kingdom of God because Jesus believed, as fellow moral perfectionist Paul did, that divine intervention was near at hand, and therefore people had to be "blameless" for only a short time (1 Thessalonians 5:23). The difficulty with perfectionism in a continuing society is evident in later traditions regarding divorce. Paul quoted Jesus' prohibition of it but then proceeded to make an exception—that if a Christian was married to an unbeliever, and the unbeliever wished a divorce, the Christian should agree to it—which he explicitly said was his own opinion, not the Lord's (1 Corinthians 7:10–16). Similarly, Matthew depicts the disciples as responding to Jesus' prohibition by proposing that if divorce is impossible it is better to avoid marriage (Matthew 19:10). The impossibility of being perfect during a full lifetime leads some modern interpreters to propose that Jesus intended these admonitions to be only an ideal, not a requirement. It is more likely, however, that Jesus the eschatological prophet regarded perfection as quite possible during the short period before the arrival of the Son of Man.

Teachings on divorce

Miracles. A prophet and teacher of ethics, Jesus was also a healer and miracle worker. In the 1st century, healers and miracle workers were fairly well known, though not precisely common, and were not considered to be superhuman beings. Jesus himself granted that others were capable of performing miracles, such as exorcisms, regardless of whether they followed him (Matthew 12:27; Mark 9:38–41; 6:7). Thus, the significance of this very important aspect of his life is frequently misunderstood. In Jesus' time, it was accepted that people could heal and perform nature miracles, such as causing rain. The question was, by what power, or spirit, they did so. Some of Jesus' opponents accused him of casting out demons by the prince of demons (Mark 3:19–22; Matthew 12:24; Luke 11:15). He countered that he did so by the spirit of God (Matthew 12:28; Luke 11:20). Obviously, many people disagreed, but this was the issue in Jesus' lifetime—not whether he, like a few others, could perform miracles, but by what power he did so. In his own day, miracles were proof neither of divinity nor of messiahship, and, at most, they might be used to validate an individual's message or way of life.

CONTROVERSY AND DANGER IN GALILEE

Crowds and autonomy. Jesus' reputation as healer had one very important historical consequence: he attracted crowds, as the early chapters of Mark (*e.g.*, 1:28, 45; 2:2) reveal. By doing so Jesus could spread his message to more people, but he also ran the risk of attracting those whose interest in him was purely selfish and who came hoping for cures only. Moreover, crowds were politically dangerous. One of the reasons Herod Antipas executed John the Baptist was because he drew such large crowds that Antipas feared an uprising (Josephus, *The Antiquities of the Jews* 18.116–119).

Although Jesus' message was not necessarily socially dangerous, the revolutionary implications of its promise of future reversal of status may have made some a little uneasy, and Jesus' promise to sinners might have been irritating to the scrupulous. Still, without crowds these aspects of his message would not have mattered very much. He did not strike at the heart of the Jewish religion as such: he did not deny the election of Abraham and the requirement of circumcision; nor did he denounce Moses and the law. Nevertheless, during his Galilean ministry some people regarded him with hostility and suspicion, partly because of the crowds and partly because of his autonomy. It was impossible to know what someone who was autonomous might do next, and this could be dangerous, especially if he had a following.

Scribes and Pharisees. In the 1st century, scribes and Pharisees were two largely distinct groups, though presumably some scribes were Pharisees. Scribes had knowledge of the law and could draft legal documents (contracts for marriage, divorce, loans, inheritance, mortgages, the sale of land, and the like). Every village had at least one scribe. Pharisees were members of a party that believed in resurrection and in following legal traditions that were ascribed not to the Bible but to "the traditions of the fathers." Like the scribes, they were also well-known legal experts: hence the partial overlap of membership of the two groups. It appears from subsequent rabbinic traditions, however, that most Pharisees were small landowners and traders, not professional scribes.

In Mark's view, Jesus' main adversaries in Galilee were scribes, but, according to Matthew, they were Pharisees. These apparently conflicting views are readily reconciled: men knowledgeable about Jewish law and tradition would have scrutinized Jesus carefully, and it is likely that both scribes and Pharisees challenged his behaviour and teaching, as the Gospels indicate (*e.g.*, Mark 2:6, 16; 3:22; Matthew 9:11; 12:2). According to one passage, the Pharisees (along with the Herodians, Mark adds) planned to destroy Jesus (Matthew 12:14; Mark 3:6). If the report of this plot is accurate, however, it seems that nothing came of it, since the Pharisees did not play a significant role in the events that led to Jesus' death. Mark and Luke assign them no role, while Matthew mentions them only once (Matthew 27:62).

Some people in Galilee may have distrusted Jesus, and legal experts probably challenged his interpretation of the law, but he was never charged formally with a serious legal offense, and opposition in Galilee did not lead to his death. Mortal danger faced Jesus only after he went to Jerusalem for what turned out to be the last time.

JESUS' LAST WEEK

Enters Jerusalem. In about the year AD 30, Jesus and his disciples went to Jerusalem from Galilee to observe Passover. Presumably they went a week early, as did tens of thousands of other Jews (perhaps as many as 200,000 or 300,000), in order to be cleansed of "corpse-impurity," in accordance with Numbers 9:10–12 and 19:1–22. The Gospels do not mention purification, but they do place Jesus near the temple in the days preceding Passover. He entered Jerusalem on a donkey, perhaps intending to recall Zechariah 9:9, which Matthew (21:5) quotes: "your king is coming to you, humble, and mounted on a donkey." This touched off a demonstration by his followers, who hailed Jesus as either "Son of David" (Matthew 21:9) or as "the one who comes in the name of the Lord" (Mark 11:9). Matthew speaks of "crowds," which suggests that many people were involved, but the demonstration was probably fairly small. Jerusalem at Passover was dangerous; it was well known to both Caiaphas, who governed the city, and Pilate, the prefect to whom the high priest was responsible, that the festivals were likely times of uprisings. Pilate's troops patrolled the roofs of the porticoes of the Temple. A large demonstration would probably have led to Jesus' immediate arrest, but, because he lived for several more days, it is likely that the crowd was relatively small.

Symbolic meaning of Jesus' entry

Jesus spent some time teaching and debating (Mark 12) and also told his disciples that the Temple would be destroyed (Mark 13:1–2). On one of the days of purification prior to the Passover sacrifice and meal, he performed his most dramatic symbolic action. He entered the part of the temple precincts where worshipers exchanged coins to pay the annual temple tax of two drachmas or bought pigeons to sacrifice for inadvertent transgressions of the law and as purificatory offerings after childbirth. Jesus turned over some of the tables (Mark 11:15–17), which led "the chief priests and the scribes" ("and the principal men of the people," Luke adds) to plan to have him executed (Mark 11:18; Luke 19:47; *cf.* Mark 14:1–2).

Later, the disciples found a room for the Passover meal, and one of them bought an animal and sacrificed it in the Temple (Mark 14:12–16; verse 16 states simply, "they prepared the passover"). Judas Iscariot, however, one of the 12, betrayed Jesus to the authorities. At the meal, Jesus blessed the bread and wine, designating the bread "my body" and the wine "my blood of the covenant" (Mark 14:22–25) or "the new covenant in my blood" (Luke 22:20 and 1 Corinthians 11:25). He also stated that he would not drink wine again until he drank it with the disciples in the kingdom (Matthew 26:29).

The Last Supper

Arrest and trials. After supper, Jesus took his disciples to the Mount of Olives to pray. While he was there, Judas led armed men sent by the chief priests to arrest him (Mark 14:43–52). They took Jesus to Caiaphas, who had gathered some of his councillors (called collectively the Sanhedrin). Jesus was first accused of threatening to destroy the Temple, but this charge was not substantiated. Caiaphas then asked him if he was "the Christ, the Son of God." According to Mark (14:61–62), Jesus said "yes" and then predicted the arrival of the Son of Man. According to Matthew (26:63–64), he said, "You say so, *but* [emphasis added] I tell you that you will see the Son of Man," apparently implying the answer was no. According to Luke he was more ambiguous: "If I tell you, you will not believe" and "You say that I am" (22:67–70). (Some scholars believe that the New International Version misrepresents Jesus' answer in Matthew and Luke.)

Whatever the answer, Caiaphas evidently had already decided that Jesus had to die. He cried "blasphemy" and rent his own garments, a dramatic sign of mourning that the Hebrew Bible prohibits the high priest from making (Leviticus 21:10). The gesture was effective, and the councillors agreed that Jesus should be sent to Pilate with the recommendation to execute him.

Charges against Jesus

It is doubtful that the titles Messiah and Son of God were actually the issue because there was no set meaning for either in 1st-century Judaism. As Mark, reprised by Matthew and Luke, presents the scene, when the attempt to have Jesus executed for threatening the Temple failed, Caiaphas simply declared whatever Jesus said (about which we must remain uncertain) to be blasphemy. This may be what convinced the council to recommend Jesus' execution. It appears, however, that the charges against Jesus that Caiaphas transmitted to Pilate (Mark 15:1–2, 26) may have included the accusation that Jesus claimed to be "king of the Jews."

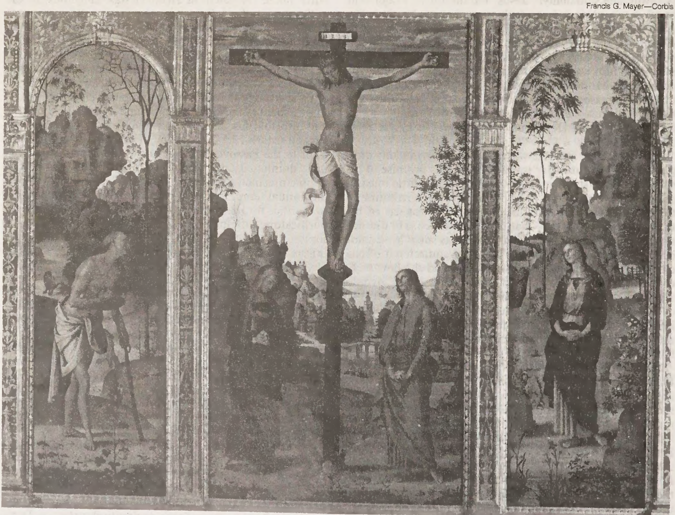
Although Pilate did not care about the fine points of Jewish law or Jesus' alleged blasphemy, most likely he saw Jesus as a potential troublemaker and therefore ordered his execution. The Gospels of Matthew, Luke, and John ascribe a rather good character to Pilate and show him as troubled over the decision but yielding to Jewish insistence (Matthew 27:11–26; Luke 21:1–25; John 18:28–40). In Luke, for example, Pilate states three times that he finds no fault with Jesus. This passage suggests that the early church, faced with making its way in the Roman Empire, did not wish its leader to be thought of as being truly guilty in Roman eyes. From other evidence Pilate is known to have been callous, cruel, and given to wanton executions (Philo, *On the Embassy to Gaius,* 300–302). He was finally dismissed from office for executing a group of Samaritans (Josephus, *The Antiquities of the Jews,* 18.85–89), and he probably sent Jesus to his death without anguishing over the decision.

Crucifixion. Crucified as would-be "king of the Jews" (Mark 15:26 and parallels Matthew 27:37; Luke 23:38; John 19:19), Jesus also was taunted on the cross as the one who would destroy and rebuild the Temple (Mark 15:29). These two charges help to explain the decision to execute him. Jesus' minor assault on the Temple and prediction of its destruction seem to be what led to his arrest. His own thinking was almost certainly that God would destroy the Temple as part of the new kingdom, perhaps rebuilding it himself (Mark 14:58). The Temple Scroll from Qumran has a similar expectation. Caiaphas and his advisers probably understood Jesus well enough: they knew that he was a prophet, not a demolition expert, and that his disciples could not damage the Temple seriously even if they were allowed to attack its walls with picks and sledges. But someone who spoke about the Temple's destruction, and who turned over tables in its precincts, was clearly dangerous. These were inflammatory acts in a city that, at festival time, was prone to uprisings that could lead to the death of many thousands of Jews. Caiaphas probably had the thought that John 11:50 attributes to him, that "it is better to have one man die for the people than to have the whole nation destroyed." The high priest, under Roman rule, was responsible for keeping the peace, and he and his advisers acted accordingly.

King of the Jews

The accusation that Jesus claimed to be "king of the Jews" was also sufficient to account for his execution. There is no direct evidence that Jesus ever said, "I am the king," but his preaching on "the kingdom of God" was inflammatory. This phrase could have been interpreted several ways, but it certainly did not mean that Rome would continue to govern Judaea. Many people resented Roman

Francis G. Mayer—Corbis



"The Crucifixion with the Virgin, Saint John, Saint Jerome, and Saint Mary Magdalene," oil on panel transferred to canvas by Perugino, c. 1485. In the National Gallery of Art, Washington, D.C.

rule, and Rome was quick to dispatch those who became too vocal in their opposition. Nevertheless, Pilate did not think that Jesus and his followers constituted a military threat. Had he thought so, he would have had the disciples, too, executed, either at the time or when they returned to Jerusalem to take up their new mission. Instead, the prefect limited his actions to their charismatic leader and turned Jesus over to his soldiers for execution. They took him and two thieves outside Jerusalem and crucified them.

Although Caiaphas did not think that Jesus could actually destroy the Temple, and Pilate did not believe that he could organize a serious revolt, inflammatory speech was a problem. Moreover, Jesus had a following, the city was packed with pilgrims who were celebrating the exodus from Egypt and Israel's liberation from foreign bondage, and Jesus had committed a small act of violence in the sacred precincts. He was dangerous, and his execution is perfectly understandable in this historical context; that is, he was executed for being what he was, an eschatological prophet. Caiaphas and his councillors fulfilled their mandate to keep the peace and suppress any signs of an uprising. Pilate presumably acted from similar motives. It is unlikely that the responsible parties lost much sleep over their decision; they were doing their duty.

Jesus' proclamation of the kingdom and his apparent threats against the Temple were based on his view that the kingdom was at hand and that he and his disciples would soon feast in it. It is possible that even to the end he expected divine intervention because among his last words was the cry "My God, my God, why have you forsaken me?" (Mark 15:34).

THE RESURRECTION

What happened next changed history in a way quite different from what Jesus seems to have anticipated. Some of his followers claimed to have seen him after his death. The details are uncertain, since the sources disagree on who saw him and where he was seen (the final sections of Matthew, Luke, and John; the beginning of Acts; and the list in Paul's first Letter to the Corinthians, 15:5–8). According to Matthew, an angel showed the empty tomb to Mary Magdalene and "the other Mary" and instructed them to tell the disciples to go to Galilee. While still in Jerusalem, the two Marys saw Jesus, who told them the same thing, and he appeared once more, to the disciples in Galilee. Matthew's account is implied in Mark 14:28 and 16:7, though the Gospel of Mark does not have a resurrection story, ending instead with the empty tomb (Mark 16:8; translations print scribal additions in brackets). According to Luke, however, while the disciples remained in Jerusalem, the women (Mary Magdalene; Joanna; Mary, the mother of James; and "the other women") found the empty tomb. "Two men in dazzling clothes" told them that Jesus had been raised. Later, Jesus appeared to two followers on the road to Emmaus (near Jerusalem), then to Peter, and later to the disciples. John (now including chapter 21, usually thought to be an appendix) mentions sightings in Galilee and Jerusalem. Acts provides a more extensive series of appearances than Luke, though written by the same author, but like it places all of these in or near Jerusalem. Paul's list of people to whom Jesus appeared does not agree very closely with the other accounts (1 Corinthians 15:5–8).

Jesus' appearances

Because of the uncertain evidence it is hard to say what really happened. Two points are important: the sources describe the resurrected Jesus as neither a resuscitated corpse, a badly wounded man staggering around, nor as a ghost. According to Luke, the first two disciples to see Jesus walked with him for several hours without recognizing him (24:13–32). Luke also reports that Jesus could disappear and reappear at will (24:31, 36). For Paul, the bodies of Christian believers will be transformed to be like the Lord's, and the resurrection body will not be "flesh and blood" (1 Corinthians 15:42–53). According to these two authors, Jesus was substantially transformed, but he was not a ghost. Luke says this explicitly (24:37–39), and Paul insists on using the word *body* as part of the term *spiritual body* rather than *spirit* or *ghost.* Luke and Paul do not agree entirely, since Luke attributes "flesh and bones" to the risen Jesus (24:39). Luke's account nevertheless requires a transformation. The authors, in other words, were trying to explain something for which they did not have a precise vocabulary, as Paul's term *spiritual body* makes clear.

It is difficult to accuse these sources, or the first believers, of deliberate fraud. A plot to foster belief in the Resurrection would probably have resulted in a more consistent story. Instead, there seems to have been a competition: "I saw him," "so did I," "the women saw him first," "no, I did; they didn't see him at all," and so on. Moreover, some of the witnesses of the Resurrection would give their lives for their belief. This also makes fraud unlikely.

The uncertainties are substantial, but, given the accounts in these sources, certainty is unobtainable. We may say of the disciples' experiences of the Resurrection approximately what the sources allow us to say of the life and message of Jesus: we have fairly good general knowledge, though many details are uncertain or dubious.

(E.P.Sa.)

The picture of Christ in the early church: The Apostles' Creed

Even before the Gospels were written, Christians were reflecting upon the meaning of what Jesus had been and what he had said and done. It is a mistake, therefore, to suppose that such reflection is a later accretion upon the simple message of the Gospels. On the contrary, the early Christian communities were engaged in witness and worship from the very beginning. The forms of that witness and worship were also the forms of the narratives in the Gospel accounts. From this fact it follows that to understand the Gospel accounts regarding Jesus we must consider the faith of the early church regarding Christ. In this sense it is valid to maintain that there is no distinction between "the Jesus of history" and "the Christ of faith," and that the only way to get at the former is by the latter. Christology, the doctrine about Christ, is then as old as Christianity itself.

To comprehend the faith of the early church regarding Christ, we must turn to the writings of the New Testament, where that faith found embodiment. It was also embodied in brief confessions or creeds, but these have not been preserved for us complete in their original form. What we have are fragments of those confessions or creeds in various books of the New Testament, snatches from them in other early Christian documents, and later forms of them in Christian theology and liturgy. The so-called Apostles' Creed is one such later form. It did not achieve its present form until quite late; just how late is a matter of controversy. But in its earliest ancestry it is very early indeed, perhaps dating back to the 1st century. And its confession regarding Christ is probably the earliest core, around which later elaborations of it were composed. Allowing for such later elaboration, we may say that in the Apostles' Creed we have a convenient summary of what the early church believed about Christ amid all the variety of its expression and formulation. The creeds were a way for Christians to explain what they meant by their acts of worship. When they put "I believe" or "We believe" at the head of what they confessed about God and Christ, they meant that their declarations rested upon faith, not merely upon observation.

PREEXISTENCE

The statement "I believe" also indicated that Christ was deserving of worship and faith, and that he was therefore on a level with God. At an early date, possibly as early as the words of Paul in Phil. 2:6–11, Christian theology began to distinguish three stages in the career of Jesus Christ: his preexistence with the Father before all things; his Incarnation and humiliation in "the days of His flesh" (Heb. 5:7), and his glorification, beginning with the Resurrection and continuing forever.

The stages of Christ's career

Probably the most celebrated statement of the preexistence of Christ is the opening verses of the Gospel of St. John. Here Christ is identified as the incarnation of the Word (Logos) through which God made all things in the beginning, a Word existing in relation to God before the creation. The sources of this doctrine have been sought in

Greek philosophy, both early and late, as well as in the Jewish thought of Philo and of the Palestinian rabbis. Whatever its source, the doctrine of the Logos in John is distinctive by virtue of the fact that it identifies the Logos with a specific historical person. Other writings of the New Testament also illustrate the faith of the early Christians regarding the preexistence of Christ. The opening chapters of both Colossians and Hebrews speak of Christ as the preexistent one through whom all things were created, therefore as distinct from the created order of things in both time and preeminence; the preposition "before" in Col. 1:17 apparently refers both to his temporal priority and to his superior dignity. Yet before any theological reflection about the nature of this preexistence had been able to find terms and concepts, the early Christians were worshipping Christ as divine. Phil. 2:6–11 may be a quotation from a hymn used in such worship. Theological reflection told them that if this worship was legitimate, he must have existed with the Father "before all ages."

Jesus Christ. By the time the text of the creed was established, this was the usual designation for the Saviour. Originally, of course, "Jesus" had been his given name, meaning "Yahweh saves," or "Yahweh will save" (see Matt. 1:21), while "Christ" was the Greek translation of the title "Messiah." Some passages of the New Testament still used "Christ" as a title (*e.g.,* Luke 24:26; II John 7), but it is evident from Paul's usage that the title became simply a proper name very early. Most of the Gentiles took it to be a proper name, and it was as "Christians" that the early believers were labelled (Acts 11:26). In the most precise language, the term "Jesus" was reserved for the earthly career of the Lord; but it seems from liturgical sources that it may actually have been endowed with greater solemnity than the name "Christ." Within a few years after the beginnings of the Christian movement, Jesus, Christ, Jesus Christ, and Christ Jesus could be used almost interchangeably, as the textual variants in the New Testament indicate. Only in modern times has it become customary to distinguish sharply among them for the sake of drawing a line between the Jesus of history and the Christ of faith, and this only in certain circles. The theologians and people of many churches still use phrases like "the life of Christ," because "Christ" is primarily a name. It is difficult to imagine how it could be otherwise when the Old Testament implications of the title have become a secondary consideration in its use—a process already evident within the New Testament.

God's only son. The declaration that Jesus Christ is the son of God is one of the most universal in the New Testament, most of whose books refer to him that way. The Gospels do not quote him as using the title for himself in so many words, although sayings like Matt. 11:27 come close to it. There are some instances where the usage of the Gospels appears to echo the more general implications of divine sonship in the Old Testament as a prerogative of Israel or of the true believer. Usually, however, it is evident that the evangelists, like Paul, meant some special honour by the name. The evangelists associated the honour with the story of Jesus' baptism (Matt. 3:17) and transfiguration (Matt. 17:5), Paul with the faith in the Resurrection (Rom. 1:4). From this association some have argued that "Son of God" in the New Testament never referred to the preexistence of Christ. But it is clear in John and in Paul that this implication was not absent, even though it was not as prominent as it became soon thereafter. What made the implication of preexistence more prominent in later Christian use of the term "Son of God" was the clarification of the doctrine of the Trinity, where "Son" was the name for the eternal Second Person (Matt. 28:19). As the Gospels show, the application of the name "Son of God" to Jesus was offensive to the Jews, probably because it seemed to smack of gentile polytheism. This also made it all too intelligible to the pagans, as early heresies indicate. Facing both the Jews and the Greeks, the apostolic church confessed that Jesus Christ was "God's only Son": the Son of God, in antithesis to Jewish claims that the eternal could have no sons; the only Son, in antithesis to Greek myths of divine procreation.

The Lord. As passages like Rom. 1:4, show, the phrase "Jesus Christ our Lord" was one of the ways the apostolic church expressed its understanding of what he had been and done. Luke even put the title into the mouth of the Christmas angel (Luke 2:11). From the way the name "Lord" (*Kyrios*) was employed during the 1st century it is possible to see several implications in the Christian use of it for Christ. The Christians meant that there were not many divine and lordly beings in the universe, but only one *Kyrios* (I Cor. 8:5–6). They meant that the Roman Caesar was not the lord of all, as he was styled by his worshippers, but that only Christ was Lord (Rev. 17:14). And they meant that Yahweh, the covenant God of the Old Testament, whose name they pronounced as "Lord," had come in Jesus Christ to establish the new covenant (see Rom. 10:12–13). Like "Son of God," therefore, the name *Kyrios* was directed against both parts of the audience to which the primitive church addressed its proclamation. At times it stood particularly for the risen and glorified Christ, as in Acts 2:36; but in passages that echoed the Old Testament it was sometimes the preexistence that was being primarily emphasized (Matt. 22:44). Gradually "our Lord," like "Christ," became a common way of speaking about Jesus Christ, even when the speaker did not intend to stress his lordship over the world.

Kyrios

INCARNATION AND HUMILIATION

Conceived by the Holy Spirit, born of the Virgin Mary. Earlier forms of the creed seem to have read: "Born of the Holy Spirit and of the Virgin Mary." The primary affirmation of this article is that the Son of God, the Word, had become man or, as John's Gospel put it, "flesh" (John 1:14). Preexistence and Incarnation presuppose each other in the Christian view of Jesus Christ. Hence the New Testament assumed his preexistence when it talked about his becoming man; and when it spoke of him as preexistent, it was ascribing this preexistence to him whom it was describing in the flesh. It may be that the reference in the creed to the Virgin Mary was intended to stress primarily her function as the guarantee of Christ's true humanity, but the creed also intended to teach the supernatural origin of that humanity. Although it is true that neither Paul nor John makes reference to it, the teaching about the virginal conception of Jesus, apparently based upon Isa. 7:14, was sufficiently widespread in the 1st century to warrant inclusion in both Matthew and Luke, as well as in creeds that date back to the 1st century. As it stands, the creedal statement is a paraphrase of Luke 1:35. In the New Testament the Holy Spirit was also involved in the baptism and the Resurrection of Jesus.

Suffered under Pontius Pilate, was crucified, dead, and buried. To a reader of the Gospels, the most striking feature of the creed is probably its omission of that which occupied a major part of the Gospels, the story of Jesus' life and teachings. In this respect there is a direct parallel between the creed and the Epistles of the New Testament, especially those of Paul. Judging by the amount of space they devoted to the Passion story, even the writers of the Gospels were apparently more interested in these few days of Jesus' life than they were in anything else he had said or done. The reason for this was the faith underlying both the New Testament and the creed, that the events of Jesus' Passion, death, and Resurrection were the events by which God had accomplished the salvation of human beings. The Gospels found their climax in those events, and the other material in them led up to those events. The Epistles applied those events to concrete situations in the early church. From the way Paul could speak of the Cross (Phil. 2:6–11) and of "the night when he [Jesus] was betrayed" (I Cor. 11:23), it seems that before our Gospels came into existence the church commemorated the happenings associated with what came to be called Holy Week. Some of the earliest Christian art was a portrayal of these happenings, another indication of their importance in the cultic and devotional life of early Christianity. How did the Cross effect the salvation of human beings? The answers of the New Testament and the early church to this question involved a variety of metaphors: Christ offered himself as a sacrifice to God; his life was a ransom for many; his death made mankind alive; his suffering was an example to people when they must suffer; he was the Second Adam, creating

The happenings of Holy Week

a new humanity; his death shows people how much God loves them; and others. Every major atonement theory of Christian theological history discussed below was anticipated by one or another of these metaphors. The New Testament employed them all to symbolize something that could be described only symbolically, that "God was in Christ reconciling the world to himself, not counting their trespasses against them" (II Cor. 5:19).

He descended into hell. This phrase was probably the last to be added to the creed. Its principal source in the New Testament was the description in I Pet. 3:18–20 of Christ's preaching to the spirits in prison. Originally the descent into hell may have been identified with the death of Christ, when he entered the abode of the dead in the underworld. But in the time before it entered the creed, the descent was frequently taken to mean that Christ had gone to rescue the souls of the Old Testament faithful from the underworld, from what western Catholic theology eventually called the *limbus patrum.* Among some of the Church Fathers the descent into hell had come to mean Christ's declaration of his triumph over the powers of hell. Despite its subsequent growth in importance, however, the doctrine of the descent into hell apparently did not form an integral part of the apostolic preaching about Christ.

GLORIFICATION

The third day he rose again from the dead. The writers of the New Testament nowhere made the Resurrection of Christ a matter for argument, but everywhere asserted it and assumed it. With it began that state in the history of Jesus Christ that was still continuing, his elevation to glory. They used it as a basis for three kinds of affirmations. The Resurrection of Christ was the way God bore witness to his son, "designated Son of God in power according to the Spirit of holiness by his resurrection from the dead" (Rom. 1:4); this theme was prominent also in the Book of Acts. The Resurrection was also the basis for the Christian hope of life after death (I Thess. 4:14), and without it that hope was said to be baseless (I Cor. 15:12–20). The Resurrection of Christ was also the ground for admonitions to manifest a "newness of life" (Rom. 6:4) and to "seek the things that are above" (Col. 3:1). The writers of the New Testament themselves expressed no doubt that the Resurrection had really happened. But Paul's discussion in I Cor. 15 shows that among those who heard the Christian message there was such doubt, as well as efforts to rationalize the Resurrection. The differences among the Gospels, and between the Gospels and Paul, suggest that from the outset a variety of traditions existed regarding the details of the Resurrection. But such differences only serve to emphasize how universal the faith in the Resurrection was amid this variety of traditions.

The Resurrection

He ascended into heaven, and sitteth on the right hand of God the father almighty. As indicated earlier, the narrative of the Ascension is peculiar to Luke-Acts, but other parts of the New Testament may refer to it. Eph. 4:8–10 may be such a reference, but many interpreters hold that for Paul Resurrection was identical with Ascension. That, they maintain, is why he could speak of the appearance of the risen Christ to him in continuity with the appearances to others (I Cor. 15:5–8) despite the fact that, in the chronology of the creed, the Ascension intervened between them. Session at the right hand of the Father was apparently a Christian interpretation of Ps. 110:1. It implied the elevation—or, as the doctrine of preexistence became clearer, the restoration—of Christ to a position of honour with God. Taken together, the Ascension and the session were a way of speaking about the presence of Christ with the Father during the interim between the Resurrection and the Second Advent. From Eph. 4:8–16, it is evident that this way of speaking was by no means inconsistent with another Christian tenet, the belief that Christ was still present in and with his church. It was, in fact, the only way to state that tenet in harmony with the doctrine of the Resurrection.

From thence he shall come to judge the quick and the dead. The creed concludes its Christological section with the doctrine of the Second Advent: the First Advent was a coming into the flesh, the Second Advent a coming in glory. Much controversy among modern scholars has been occasioned by the role of this doctrine in the early church. Those who maintain that Jesus erroneously expected the early end of the world have often interpreted Paul as the first of those who began the adjustment to a delay in the end, with John's Gospel as a more advanced stage of that adjustment. Those who hold that the imminence of the end was a continuing aspect of human history as Jesus saw it also maintain that this phrase of the creed was a statement of that imminence, without any timetable necessarily implied. From the New Testament it seems that both the hope of the Second Coming and a faith in the continuing presence of Christ belonged to the outlook of the apostolic church, and that seems to be what the creed meant. The phrase "the quick and the dead" is a summary of passages like I Cor. 15:51–52 and I Thess. 4:15–17.

The Second Advent

In order to complete the confession of the creed regarding the glorification of Christ, the Nicene Creed added the phrase: "Of His kingdom there shall be no end." This was a declaration that Christ's return as judge would usher in the full exercise of his reign over the world. Such was the expectation of the apostolic church, based upon what it knew and believed about Jesus Christ.

The dogma of Christ in the ancient councils

The main lines of orthodox Christian teaching about the person of Christ were set by the New Testament and the ancient creeds. But what was present there in a germinal form became a clear statement of Christian doctrine when it was formulated as dogma. In one way or another, the first four ecumenical councils were all concerned with the formulation of the dogma regarding the person of Christ—his relation to the Father, and the relation of the divine and the human in him.

Such a formulation became necessary because teachings arose within the Christian community that seemed to threaten what the church believed and confessed about Christ. Both the dogma and the heretical teachings against which the dogma was directed are therefore part of the history of Jesus Christ.

THE COUNCILS OF NICAEA AND CONSTANTINOPLE

Early heresies. From the outset, Christianity has had to contend with those who misinterpreted the person and mission of Jesus. Both the New Testament and the early confessions of the church referred and replied to such misinterpretations. As the Christian movement gained adherents from the non-Jewish world, it had to explain Christ in the face of new challenges.

These misinterpretations touched both the question of his humanity and the matter of his deity. A concern to safeguard the true humanity of Jesus led some early Christians to teach that Jesus of Nazareth, an ordinary man, was adopted as the Son of God in the moment of his baptism or after his Resurrection; this heresy was called adoptionism. Gnostics and others wanted to protect him against involvement in the world of matter, which they regarded as essentially evil, and therefore taught that he had only an apparent, not a real body; they were called docetists. Most of the struggle over the person of Christ, however, dealt with the question of his relation to the Father. Some early views were so intent upon asserting his identity with the Father that the distinction of his person was lost and he became merely a manifestation of the one God. Because of this idea of Christ as a "mode" of divine self-manifestation, proponents of this view were dubbed "modalists"; from an early supporter of the view it was called "Sabellianism." Other interpretations of the person of Christ in relation to God went to the opposite extreme. They insisted so strenuously upon the distinctness of his person from that of the Father that they subordinated him to the Father. Many early exponents of the doctrine of the Logos were also subordinationists, so that the Logos idea itself became suspect in some quarters. What was needed was a framework of concepts with which to articulate the doctrine of Christ's oneness with the Father and yet distinctness from the Father, and thus to answer the question (Adolf von Harnack): "Is the Divinity which has appeared on earth and reunited men with god identical with that supreme Divinity which governs heaven and earth, or is it a demigod?"

Nicaea. That question forced itself upon the church through the teachings of Arius. He maintained that the Logos was the first of the creatures, called into being by God as the agent or instrument through which he was to make all things. Christ was thus less than God, but more than man; he was divine, but he was not God. To meet the challenge of Arianism, which threatened to split the church, the newly converted emperor Constantine convoked in 325 the first ecumenical council of the Christian church at Nicaea. The private opinions of the attending bishops were anything but unanimous, but the opinion that carried the day was that espoused by the young presbyter Athanasius, who later became bishop of Alexandria. The Council of Nicaea determined that Christ was "begotten, not made," that he was therefore not creature but creator. It also asserted that he was "of the same essence as the Father" (*homoousios to patri*). In this way it made clear its basic opposition to subordinationism, even though there could be, and were, quarrels about details. It was not equally clear how the position of Nicaea and of Athanasius differed from modalism. Athanasius asserted that it was not the Father nor the Holy Spirit, but only the Son that became incarnate as Jesus Christ. But in order to assert this, he needed a more adequate terminology concerning the persons in the Holy Trinity. So the settlement at Nicaea regarding the person of Christ made necessary a fuller clarification of the doctrine of the Trinity, and that clarification in turn made possible a fuller statement of the doctrine of the person of Christ.

The first ecumenical council

Constantinople. Nicaea did not put an end to the controversies but only gave the parties a new rallying point. Doctrinal debate was complicated by the rivalry among bishops and theologians as well as by the intrusion of imperial politics that had begun at Nicaea. Out of the post-Nicene controversies came that fuller statement of the doctrine of the Trinity which was needed to protect the Nicene formula against the charge of failing to distinguish adequately between the Father and the Son. Ratified at the Council of Constantinople in 381, but since lost, that statement apparently made official the terminology developed by the supporters of Nicene orthodoxy in the middle of the 4th century: one divine essence, three divine persons (*mia ousia, treis hypostaseis*). The three persons, Father, Son, and Holy Spirit, were distinct from one another but were equal in their eternity and power. Now it was possible to teach, as Nicaea had, that Christ was "of the same essence as the Father" without arousing the suspicion of modalism. Although this doctrine seemed to make problematical the unity of God, it did provide an answer to the first of the two issues confronted by the church in its doctrine of the person of Christ—the issue of Christ's relation to the Father. It now became necessary to clarify the second issue—the relation of the divine and the human within Christ.

The Trinity

THE COUNCILS OF EPHESUS AND CHALCEDON

By excluding several extreme positions from the circle of orthodoxy, the formulation of the doctrine of the Trinity in the 4th century determined the course of subsequent discussion about the person of Christ. It also provided the terminology for that discussion, since 5th-century theologians were able to describe the relation between the divine and the human Christ by analogy to the relation between the Father and the Son in the Trinity. The term that was found to express this relation in Christ was "nature," *physis*. There were three divine persons in one divine essence; such was the outcome of the controversies in the 4th century. But there were also two natures, one of them divine and the other human, in the one person Jesus Christ. Over the relation between these two natures the theologians of the 5th century carried on their controversy.

The abstract questions with which they sometimes dealt in that controversy, some of them almost unintelligible to a modern mind, must not be permitted to obscure the fact that a basic issue of the Christian faith was at stake: how can Jesus Christ be said to partake of both identity with God and brotherhood with humanity?

The parties. During the half century after the Council of Constantinople several major points of emphasis developed in the doctrine of the person of Christ; characteristically, these are usually defined by the episcopal see that espoused them. There was a way of talking about Christ that was characteristic of the see at Alexandria. It stressed the divine character of all that Jesus Christ had been and done, but its enemies accused it of absorbing the humanity of Christ in his divinity. The mode of thought and language employed at Antioch, on the other hand, emphasized the true humanity of Christ; but its opponents maintained that in so doing it had split Christ into two persons, each of whom maintained his individual selfhood while they acted in concert with each other. Western theology was not as abstract as either of these alternatives. Its central emphasis was a practical concern for human salvation and for as irenic a settlement of the conflict as was possible without sacrificing that concern. Even more than in the 4th century, considerations of imperial politics were always involved in conciliar actions, together with the fear in countries like Egypt that Constantinople might come to dominate them. Thus a decision regarding the relation between the divine and the human in Christ could be simultaneously a decision regarding the political situation. Nevertheless, the settlements at which the councils of the 5th century arrived may be and are regarded as normative in the church long after their political setting has disappeared.

The conflict between Alexandria and Antioch came to a head when Nestorius, taking exception to the use of the title "Mother of God" or, more literally, "God-Bearer" (*Theotokos*) for the Virgin Mary, insisted that she was only "Christ-Bearer." In this insistence the Antiochian emphasis upon the distinction between the two natures in Christ made itself heard throughout the church. The Alexandrian theologians responded by charging that Nestorius was dividing the person of Christ, which they represented as so completely united that, in the famous phrase of Cyril, there was "one nature of the Logos which became incarnate." By this he meant that there was only one nature, the divine, before the Incarnation, but that after the Incarnation there were two natures indissolubly joined in one person; Christ's human nature had never had an independent existence. There were times when Cyril appeared to be saying that there was "one nature of the incarnate Logos" even after the Incarnation, but his most precise formulations avoided this language.

The Council of Ephesus in 431 was one in a series of gatherings called to settle this conflict, some by one party and some by the other. The actual settlement was not accomplished, however, until the calling of the Council of Chalcedon in 451.

The settlement at Chalcedon. The basis of the settlement was the Western understanding of the two natures in Christ, as formulated in the *Tome* of Pope Leo I of Rome. Chalcedon declared: "We all unanimously teach . . . one and the same Son, our Lord Jesus Christ, perfect in deity and perfect in humanity . . . in two natures, without being mixed, transmuted, divided, or separated. The distinction between the natures is by no means done away with through the union, but rather the identity of each nature is preserved and concurs into one person and being." In this formula the valid emphases of both Alexandria and Antioch came to expression; both the unity of the person and the distinctness of the natures were affirmed. Therefore the decision of the Council of Chalcedon has been the basic statement of the doctrine of the person of Christ for most of the church ever since. The western part of the church went on to give further attention to the doctrine of the work of Christ. In the eastern part of the church the Alexandrians and the Antiochians continued the controversies that had preceded Chalcedon, but they clashed now over the question of how to interpret Chalcedon. The controversy over the Monophysite and the Monothelite heresies was an effort to clarify the interpretation of Chalcedon, with the result that the extremes of the Alexandrian position were condemned just as the Nestorian extreme of the Antiochian had been.

The two natures in Christ

Emerging from all this theological discussion was an interpretation of the person of Christ that affirmed both his oneness with God and his oneness with humanity while still maintaining the oneness of his person. Interestingly, the liturgies of the church had maintained this interpretation at a time when the theologians of the church were still struggling for clarity; and the final solution was a scientifically precise restatement of what had been present germinally in the liturgical piety of the church. In the formula of Chalcedon that solution finally found the framework of concepts and of vocabulary that it needed to become intellectually consistent. In one sense, therefore, what Chalcedon formulated was what Christians had been believing from the beginning; but in another sense it represented a development from the earlier stages of Christian thought.

Adam Woolfitt—Corbis

Detail of Jesus Christ from the Deeis Mosaic, 12th century. From the Hagia Sophia, Istanbul, Turkey.

The interpretation of Christ in Western faith and thought

With the determination of the orthodox teaching of the church regarding the person of Christ, it still remained necessary to clarify the doctrine of the work of Christ. While it had been principally in the East that the discussion of the former question was carried on—though with important additions from the West, as we have seen—it was the Western Church that provided the most detailed answers to the question: granted that this is what Jesus Christ was, how are we to describe what it is that he did?

DOCTRINES OF THE PERSON AND WORK OF CHRIST

St. Augustine

The medieval development. The most representative spokesman of the Western Church on this question, as on most others, was St. Augustine. His deep understanding of the meaning of human sin was matched by his detailed attention to the meaning of divine grace. Central to that attention was his emphasis upon the humanity of Jesus Christ as man's assurance of his salvation, an emphasis to which he gave voice in a variety of ways. The humanity of Christ showed how God elevated the humble; it was the link between the physical nature of human beings and the spiritual nature of God; it was the sacrifice which the human race offered to God; it was the foundation of a new humanity, recreated in Christ as the old humanity had been created in Adam—in these and other ways Augustine sought to describe the importance of the Incarnation for the redemption of man. By combining this stress upon the humanity of Christ as the Saviour with a doctrine of the Trinity that was orthodox but nevertheless highly creative and original, St. Augustine put his mark indelibly upon Western piety and theology, which, in Anselm and in the reformers, sought further for adequate language in which to describe God's deed of reconciliation in Jesus Christ.

During the formative centuries of Christian dogma, there had been many ways of describing that reconciliation, most of them having some precedent in biblical speech. One of the most prominent pictures of the reconciliation was that connected with the biblical metaphor of ransom: Satan held the human race captive in its sin and corruptibility, and the death of Christ was the ransom paid to the Devil as the price for setting mankind free. A related metaphor was that of the victory of Christ: Christ entered into mortal combat with Satan for the human race, and the winner was to be lord; although the Crucifixion appeared to be Christ's capitulation to the enemy, his Resurrection broke the power of the Devil and gave the victory to Christ, granting to mankind gift of immortality. From the Old Testament and the Epistle to the Hebrews came the image of Christ as the sacrificial victim who was offered up to God as a means of stilling the divine anger. From the sacrament of penance came the idea, most fully developed by St. Anselm, that the death of Christ was a vicarious satisfaction rendered for mankind. Like the New Testament, the Church Fathers could admonish their hearers to learn from the death of Christ how to suffer patiently. They could also point to the suffering and death of Christ as the supreme illustration of how much God loves mankind. As in the New Testament, therefore, so in the tradition of the church there were many figures of speech to represent the miracle of the reunion between man and God effected in the God-man Christ Jesus.

Common to all these figures of speech was the desire to do two things simultaneously: to emphasize that the reunion was an act of God, and to safeguard the participation of man in that act. Some theories were so "objective" in their emphasis upon the divine initiative that man seemed to be almost a pawn in the transaction between God in Christ and the Devil. Other theories so "subjectively" concentrated their attention upon man's involvement and man's response that the full scope of the redemption could vanish from sight. It was in Anselm of Canterbury that Western Christendom found a theologian who could bring together elements from many theories into one doctrine of the Atonement, summarized in his book, *Cur Deus homo?* According to this doctrine, sin was a violation of the honour of God. God offered man life if he rendered satisfaction for that violation; but the longer man lived, the worse the situation became. Only a life that was truly human and yet had infinite worth would have been enough to give such a satisfaction to the violated honour of God on behalf of the entire human race. Such a life was that of Jesus Christ, whom the mercy of God sent as a means of satisfying the justice of God. Because he was true man, his life and death could be valid for men; because he was true God, his life and death could be valid for *all* men. By accepting the fruits of his life and death, mankind could receive the benefits of his satisfaction. With some relatively minor alterations, Anselm's doctrine of Atonement passed over into the theology of the Latin church, forming the basis of both Roman Catholic and orthodox Protestant ideas of the work of Christ. It owed its acceptance to many factors, not the least of them being the way it squared with the liturgy and art of the West. The crucifix has become the traditional symbol of Christ in the Western Church, reinforcing and being reinforced by the satisfaction theory of the Atonement.

Scholastic theology

Scholastic theology, therefore, did not modify traditional ways of speaking about either the person or the work of Christ as sharply as it did, for example, some of the ways the Church Fathers had spoken about the presence of the body and blood of Christ in the Eucharist. The major contribution of the scholastic period to the Christian conception of Jesus Christ appears to lie in the way it managed to combine theological and mystical elements. Alongside the growth of Christological dogma and sometimes in apparent competition with it was the development of a view of Christ that sought personal union with him rather than accurate concepts about him. Such a view of Christ appeared occasionally in the writings of Augustine, but it was in men like Bernard of Clairvaux that it attained both its fullest expression and its most adequate harmonization with the dogmatic view. The relation between the divine and the human natures in Christ, as formulated in ancient dogma, provided the mystic with the ladder he needed to ascend through the man Jesus to the eternal Son of God, and through him to a mystical union with the Holy Trinity; this had been anticipated in the mystical theology of some of the Greek fathers. At the same time the dogma saved mysticism from the pantheistic excesses to which it might otherwise have gone; for the doctrine of the two natures meant that the humanity of the Lord was not an expendable element in Christian piety, mystical or not, but its indispensable presupposition and the continuing object of its adoration, in union with his deity. As a matter of fact, another contribution of the medieval development was the increased emphasis of St. Francis of Assisi and his followers upon the human life of Jesus. These brotherhoods cultivated a more practical and ethical version of mystical devotion, to be distinguished from speculative and contemplative mysticism. Their theme became the imitation of Christ in a life of humility and obedience. With it came a new appreciation of that true humanity of Christ which the dogma had indeed affirmed, but which theologians had been in danger of reducing to a mere dogmatic concept. As Henry Thode and others have suggested, this new appreciation is reflected in the way painters like Giotto began to portray Jesus, in contrast with their Western predecessors and especially with the stylized picture of Christ in Byzantine icon painting.

Luther, Calvin, and Zwingli

The Reformation and classical Protestantism. The attitude of the reformers toward traditional conceptions of the person and work of Christ was conservative. Insisting for both religious and political reasons that they were orthodox, they altered little in the Christological dogma. Martin Luther and John Calvin gave the dogma a new meaning when they related it to their doctrine of justification by grace through faith. Because of his interpretation of sin as the captivity of the will, Luther also revived the patristic metaphor of the Atonement as the victory of Christ; it is characteristic of him that he wrote hymns for both Christmas and Easter but not for Lent. The new attention to the Bible that came with the Reformation created interest in the earthly life of Jesus, while the Reformation idea of "grace alone" and of the sovereignty of God even in his grace made the deity of Christ a matter of continuing importance.

In the ideas about the Lord's Supper set forth by Huldrych Zwingli, Luther thought he saw a threat to the orthodox doctrine of Christ, and he denounced those doctrines vehemently. As this controversy progressed, Luther interpreted the ancient dogma of the two natures to mean that the omnipresence of the divine nature was communicated to the human nature of Christ, and that therefore Christ as both God and man was present everywhere and at all times. Although he repudiated both Luther's and Zwingli's theories, Calvin was persuaded that the ancient Christological dogma was true to the biblical witness and he permitted no deviation from it. All this is evidence for the significance that "Jesus Christ, true God begotten of the Father from eternity, and also true man, born of the Virgin Mary," to use Luther's formula, had in the faith and theology of all the reformers.

At one point the theology of the reformers did serve to bring together several facets of the biblical and the patristic descriptions of Jesus Christ. That was the doctrine of the threefold office of Christ, systematized by Calvin and developed more fully in Protestant orthodoxy: Christ as prophet, priest, and king. Each of these symbolized the fulfillment of the Old Testament and represented one aspect of the church's continuing life. Christ as prophet fulfilled and elevated the prophetic tradition of the Old Testament, while continuing to fulfill his prophetic office in the ministry of the Word. Christ as priest brought to an end the sacrificial system of the Old Testament by being both the priest and the victim, while he continues to function as intercessor with and for the church. Christ as king was the royal figure to whom the Old Testament had pointed, while exercising his rule among men now through those whom he has appointed. In each of the three, Protestants differed from one another according to their theological, ethical, or liturgical positions. But the threefold office enabled Protestant theology to take into account the complexity of the biblical and patristic pictures of Christ as no oversimplified theory was able to do, and it is probably the chief contribution of the reformers to the theological formulation of the doctrine of the person and work of Christ.

THE DEBATE OVER CHRISTOLOGY IN MODERN CHRISTIAN THOUGHT

Few Protestant theologians in the middle of the 20th century were willing to endorse the ancient dogma of the two natures in Christ as unconditionally as the reformers had done, for between the Reformation and modern theology there intervened a debate over Christology that altered the perspective of most Protestant denominations and theologians. By the 20th century there was a wider gap between the theology of the reformers and that of many modern Protestants than there had been between the theology of the reformers and that of their Roman Catholic opponents.

The Socinian criticism

Origins of the debate. The earliest criticism of orthodox dogma came in the age of the Reformation, not from the reformers but from the "left wing of the Reformation," from Michael Servetus (1511?–53) and the Socinians. This criticism was directed against the presence of nonbiblical concepts and terms in the dogma, and it was intent upon safeguarding the true humanity of Jesus as a moral example. There were many inconsistencies in this criticism, such as the willingness of Servetus to call Jesus "Son of God" and the Socinian custom of addressing prayer and worship to him. But it illustrates the tendency, which became more evident in the Enlightenment, to use the Reformation protest against Catholicism as a basis for a protest against orthodox dogma as well. While that tendency did not gain much support in the 16th century because of the orthodoxy of the reformers, later criticism of orthodox Christology was able to wield the "Protestant principle" against the dogma of the two natures on the grounds that this was a consistent application of what the reformers had done. Among the ranks of the Protestant laity, the hymnody and the catechetical instruction of the Protestant churches assured continuing support for the orthodox dogma. Indeed, the doctrine of Atonement by the vicarious satisfaction of Christ's death has seldom been expressed as amply as it was in the hymns and catechisms of both the Lutheran and the Reformed churches. During the period of Pietism in the Protestant churches, this loyalty to orthodox teaching was combined with a growing emphasis upon the humanity of Jesus, also expressed in the hymnody of the time.

When theologians began to criticize orthodox ideas of the person and work of Christ, therefore, they met with opposition from the common people. Albert Schweitzer dates the development of a critical attitude from the work of H.S. Reimarus (1694–1768), but Reimarus was representative of the way the Enlightenment treated the traditional view of Jesus. The books of the Bible were to be studied just as other books are, and the life of Jesus was to be drawn from them by critically sifting and weighing the evidence of the Gospels. The Enlightenment thus initiated the modern interest in the life of Jesus, with its detailed attention to the problem of the relative credibility of the Gospel records. It has been suggested by some historians that the principal target of Enlightenment criticism was not the dogma of the two natures but the doctrine of the vicarious Atonement. The leaders of Enlightenment thought did not make a sudden break with traditional ideas, but gave up belief in miracles, the Virgin Birth, the Resurrection, and the Second Advent only gradually. Their principal importance for the history of the doctrine of Christ consists in the fact that they made the historical study of the sources for the life of Jesus an indispensable element of any Christology.

The 19th century. Although the Enlightenment of the 18th century was the beginning of the break with orthodox teachings about Jesus Christ, it was only in the 19th century that this break attracted wide support among theologians and scholars in many parts of Christendom—even, for a while, among the Modernists of the Roman Catholic Church. Two works of the 19th century were especially influential in their rejection of orthodox Christology. One was the *Life of Jesus*, first published in 1835 by David Friedrich Strauss; the other, bearing the same title, was first published by Ernest Renan in 1863. Strauss's work paid more attention to the growth of Christian ideas—he called them "myths"—about Jesus as the basis for the picture we have in the Gospels, while Renan attempted to account for Jesus' career by a study of his inner psychological life in relation to his environment. Both works achieved wide circulation and were translated into other languages, including English. They took up the Enlightenment contention that the sources for the life of Jesus were to be studied as other sources are, and what they constructed on the basis of the sources was a type of biography in the modern sense of the word. In addition to Strauss and Renan, the 19th century saw the publication of a plethora of books about the life and teachings of Jesus. Each new hypothesis regarding the problem of the Synoptic Gospels implied a reconstruction of the life and message of Jesus.

Strauss and Renan

The fundamental assumption for most of this work on the life and teachings of Jesus was a distinction between the "Jesus of history" and the "Christ of faith." Another favourite way of putting the distinction was to speak of the religion *of* Jesus in antithesis to the religion *about* Jesus. This implied that Jesus was a man like other men, but with a heightened awareness of the presence and power of God. Then the dogma of the church had mistaken this awareness for a metaphysical statement that Jesus was the Son of God and had thus distorted the original simplicity of his message. Some critics went so far as to question the very historicity of Jesus, but even those who did not go that far questioned the historicity of some of the sayings and deeds attributed to Jesus in the Gospels.

In part this effort grew out of the general concern of 19th-century scholarsip with the problem of history, but it also reflected the religious and ethical assumptions of the theologians. Many of them were influenced by the moral theories of Kant in their estimate of what was permanent about the teachings of Jesus, and by the historical theories of Hegel in the way they related the original message of Jesus to the Christian interpretations of that message by later generations of Christians. The ideas of evolution and of natural causality associated with the science of the 19th century also played a part through the naturalistic explanations of the biblical miracles. And the historians of dogma, climaxing in Adolf von Harnack (1851–1931), used their demonstration of the dependence of ancient Christology upon non-Christian sources for its concepts and terminology to reinforce their claim that Christianity had to get back from the Christ of dogma to the "essence of Christianity" in the teachings of Jesus about the fatherhood of God and the brotherhood of man.

The 20th century. At the beginning of the 20th century the most influential authorities on the New Testament were engaged in this quest for the essence of Christianity and for the Jesus of history. But that quest led in the early decades of the 20th century to a revolutionary conclusion regarding the teachings of Jesus, namely, that he had expected the end of the age to come shortly after his death and that his teachings as laid down in the Gospels were an "interim ethic," intended for the messianic community in the brief span of time still remaining before the end. The effort to apply those teachings in modern life was criticized as a dangerous modernization. This thesis of the "consistent eschatology" in Jesus' message was espoused by Johannes von Weiss (1863–1914) and gained wide circulation through the writings of Albert Schweitzer.

The years surrounding World War I also saw the development of a new theory regarding the composition of the Gospels. Because of its origin, this theory is usually called form criticism (German *Formgeschichte*). It stressed the forms of the Gospel narratives—parables, sayings, miracle stories, Passion accounts, etc.—as an indication of the oral tradition in the Christian community out of which the narratives came. While the attention of earlier scholars had been concentrated on the authenticity of Jesus' teachings

Form criticism

Jim Zuckerman—Corbis

Christ the Redeemer statue on Mount Corcovado, overlooking Rio de Janeiro, Brazil.

as transmitted in the Gospels, this new theory was less confident of being able to separate the authentic from the later elements in the Gospel records, though various proponents of it did suggest criteria by which such a separation might be guided. The studies of form criticism made a life of Jesus in the old biographical sense impossible, just as consistent eschatology had declared impossible the codification of a universal ethic from the teachings of Jesus. Some adherents of form criticism espoused an extreme skepticism regarding any historical knowledge of Jesus' life at all, but the work of men like Martin Dibelius and even Rudolf Bultmann showed that such skepticism was not warranted by the conclusions of this study.

Influenced by these trends in New Testament study, Protestant theology by the middle of the 20th century was engaged in a reinterpretation of the Christology of the early church. Some Protestant churches continued to repeat the formulas of ancient dogma, but even there the critical study of the New Testament documents was beginning to call those formulas into question. The struggles of the evangelical churches in Germany under Adolf Hitler caused some theologians to realize anew the power of the ancient dogma of the person of Christ to sustain faith, and some of them were inclined to treat the dogma with less severity. But even they acknowledged that the formulation of that dogma in static categories of person, essence, and nature was inadequate to the biblical emphasis upon actions and events rather than upon states of being. Karl Barth for the Reformed tradition, Lionel Thornton for the Anglican tradition, and Karl Heim for the Lutheran tradition were instances of theologians trying to reinterpret classical Christology. While yielding nothing of their loyalty to the dogma of the church, Roman Catholic theologians like Karl Adam were also endeavouring to state that dogma in a form that was meaningful to modern men. The doctrine of the work of Christ was receiving less attention than the doctrine of Christ's person. In much of Protestantism, the concentration of the 19th century upon the teachings of Jesus had made it difficult to speak of more than the prophetic office. The priestly office received least attention of all; and, therefore, despite the support accorded to efforts like that of Gustaf Aulén to reinterpret the metaphor of the Atonement as Christ's victory over his enemies, Protestant theology in the middle of the 20th century was still searching for a doctrine of the Atonement to match its newly won insights into the doctrine of the person of Christ.

In a curious way, therefore, the figure of Jesus Christ has become both a unitive and a divisive element in Christendom. All Christians are united in their loyalty to him, even though they express their loyalty in a variety of doctrinal and liturgical ways. But doctrine and liturgy also divide Christian communions from one another. It has not been the official statements about Christ that have differed widely among most communions. What has become a sharp point of division is the amount of historical and critical inquiry that is permitted where the person of Christ is involved. Despite their official statements and confessions, most Protestant denominations had indicated by the second half of the 20th century that they would tolerate such inquiry, differ though they did in prescribing how far it would be permitted to go. On the other hand, the exclusion of Modernism by the Roman Catholic Church in 1907–10 drew definite limits beyond which the theological use of the methods of critical inquiry was heretical. Within those limits, however, Roman Catholic biblical scholars were engaging in considerable critical literary study, at the same time that critical Protestant theologians were becoming more sympathetic to traditional Christological formulas.

(J.J.Pe.)

BIBLIOGRAPHY

Texts, translations, and synopses: The principal source for the life and teachings of Jesus is the Greek New Testament, and the most useful edition is EBERHARD NESTLE *et al.* (eds.), *Novum Testamentum Graece,* 26th new rev. ed. (1979). The best edition of the synopsis of the first three Gospels, which prints the text in parallel columns, is ALBERT HUCK, HANS LIETZMANN, and FRANK LESLIE CROSS (eds.), *A Synopsis of the First Three Gospels,* 9th ed. (1936, reissued 1976); and the best translation of Huck is BURTON H. THROCKMORTON, JR. (ed.), *Gospel Parallels: A Comparison of the Synoptic Gospels,* 5th ed. (1992). E.P. SANDERS and MARGARET DAVIES, *Studying the Synoptic Gospels* (1989), examines the difficulties of using the Synoptic Gospels. There are also numerous translations of the Bible, and among the better are WAYNE A. MEEKS *et al.* (eds.), *The HarperCollins Study Bible: New Revised Standard Version, with the Apocryphal/Deuterocanonical Books* (1993); and HERBERT G. MAY and BRUCE M. METZGER (eds.), *The New Oxford Annotated Bible with the Apocrypha: Revised Standard Version, Containing the Second Edition of the New Testament and an Expanded Edition of the Apocrypha* (1977).

There are several noncanonical sources that may shed light on the life and teachings of Jesus. Good editions of the noncanonical sources are EDGAR HENNECKE (ed.), *New Testament Apocrypha,* 2 vol., ed. by WILHELM SCHNEEMELCHER, vol. 1, *Gospels and Related Writings,* English translation ed. by R.MCL. WILSON (1963; trans. from German 3rd ed., 1959–64); and BENTLEY LAYTON (ed.), *Nag Hammadi Codex II, 2–7,* 2 vol. (1989), for the Gospel of Thomas. Jewish sources that provide the context in which Jesus lived and worked include JAMES H. CHARLESWORTH (ed.), *The Old Testament Pseudepigrapha,* vol. 1, *Apocalyptic Literature and Testaments* (1983); FLAVIUS JOSEPHUS, *Josephus,* trans. by H.ST.J. THACKERY, RALPH MARCUS, and L.H. FELDMAN, 9 vol. (1926–65, reissued in 10 vol., 1981–93), text in Greek and English; PHILO, *Philo: With an English Translation,* vol. 10, *On the Embassy to Gaius,* trans. from Greek by F.H. COLSON (1962, reprinted 1998); HERBERT DANBY (ed. and trans.), *The Mishnah* (1933), available in many later printings; and GEZA VERMES (trans. and ed.), *The Complete Dead Sea Scrolls in English* (1997). Important Roman sources include SUETONIUS, *Suetonius,* rev. ed., trans. from Latin by J.C. ROLFE, vol. 2, *Lives of the Caesars, Book V–VIII* (1998), text in Latin and English; and MOSES HADAS (ed.), *The Complete Works of Tacitus,* trans. from Latin by ALFRED JOHN CHURCH and WILLIAM JACKSON BRODRIBB (1942).

General books on the historical Jesus: ALBERT SCHWEITZER, *The Quest of the Historical Jesus* (1910, reissued 1998; originally published in German, 1906), establishes the importance of Jewish eschatology for understanding Jesus and his teachings. RUDOLF BULTMANN, *Jesus and the Word* (1934, reissued 1989; originally published in German, 1926), accepts the importance of eschatology but interprets it in existentialist categories. C.H. DODD, *The Founder of Christianity* (1970, reissued 1973), redefines eschatology as referring not to the future but to a higher order. GEZA VERMES, *Jesus the Jew: A Historian's Reading of the Gospels,* 2nd ed. (1983), describes Jesus as thoroughly Jewish and as a Galilean charismatic, and *The Gospel of Jesus the Jew* (1981) examines the Jewish parallels to Jesus' teaching. BEN F. MEYER, *The Aims of Jesus* (1979), proposed, against the then prevailing opinion, that Jesus' intentions can be discerned. E.P. SANDERS, *Jesus and Judaism,* 2nd ed. (1987), and *The Historical Figure of Jesus* (1993, reissued 1996), place Jesus in the context of Palestinian Judaism and reject excessive historical skepticism about the life of Jesus. JOHN DOMINIC CROSSAN, *The Historical Jesus: The Life of a Mediterranean Jewish Peasant* (1991), and *Jesus: A Revolutionary Biography* (1994), argue that Jesus was not a Jewish eschatological prophet but a cynic-like philosopher and that Galilee was Hellenistic and not distinctively Jewish. This is an extremely eccentric interpretation. JOHN P. MEIER, *A Marginal Jew: Rethinking the Historical Jesus,* 2 vol. (1991–94), is a very thorough account along the lines established by Schweitzer, Vermes, and Sanders.

Along with the general works, there are many studies of specific aspects of the life and teachings of Jesus. RAYMOND E. BROWN, *The Birth of the Messiah: A Commentary on the Infancy Narratives in Matthew and Luke,* new updated ed. (1993, reissued 1999), is an encyclopedic account of the birth narratives in the Gospels. MORTON SMITH, *Jesus the Magician* (1978, reissued 1998), is a study of the miracles of Jesus and their parallels in Hellenistic traditions. RAYMOND E. BROWN, *The Death of the Messiah, from Gethsemane to the Grave: A Commentary on the Passion Narratives in the Four Gospels,* 2 vol. (1994, reissued 1998), is an excellent introduction to the Passion of Jesus. PAULA FREDRIKSEN, *Jesus of Nazareth, King of the Jews: A Jewish Life and the Emergence of Christianity* (1999), emphasizes the importance of the Crucifixion for understanding the life and teachings of Jesus.

First-century Judaism: The standard account of all aspects of 1st-century Judaism, though it is weak on the Jewish religion, is EMIL SCHÜRER, *The History of the Jewish People in the Age of Jesus Christ (175 B.C.–A.D. 135),* trans. from German, rev. and ed. by GEZA VERMES and FERGUS MILLAR, 3 vol. in 4 (1973–87). E.P. SANDERS, *Judaism: Practice and Belief, 63 BCE–66 CE* (1992, reprinted with corrections 1994), provides a good overview of the Jewish religion in the 1st century. GEZA VERMES, *An Introduction to the Complete Dead Sea Scrolls,* rev. ed. (2000), is a good study of the collection at Qumrān. YIGAEL YADIN (ed.), *The Temple Scroll,* 3 vol. in 4 (1977–83; originally published in Hebrew, 1977), is a valuable study of the longest and most important scroll from Qumrān.

The Roman Empire in the East: FERGUS MILLAR, *The Roman Near East: 31 B.C.–A.D. 337* (1993, reprinted 1996); and A.H.M. JONES, *The Cities of the Eastern Roman Provinces*, 2nd ed., rev. by MICHAEL AVI-YONAH *et al.* (1971, reprinted 1983), are the standard works on the Roman Empire in the eastern Mediterranean.

Galilee, Tiberias, and Sepphoris: SEÁN FREYNE, *Galilee, from Alexander the Great to Hadrian, 323 B.C.E. to 135 C.E.: A Study of Second Temple Judaism* (1980, reissued 1998), is the standard study of the ancient Galilee. SEÁN FREYNE, *Galilee, Jesus, and the Gospels: Literary Approaches and Historical Investigations* (1988), and LEE I. LEVINE (ed.), *The Galilee in Late Antiquity* (1992), are also good studies of the region in antiquity. A.H.M. JONES, *The Greek City from Alexander to Justinian* (1940, reissued 1979), is a good introduction to ancient cities like Tiberias and Sepphoris. ZEEV WEISS, "Sepphoris," in EPHRAIM STERN, AYELET LEWINSON-GILBOA, and JOSEPH AVIRAM (eds.), *The New Encyclopedia of Archaeological Excavations in the Holy Land*, 4 vol. (1993; originally published in Hebrew, 1992), pp. 1324–28; YIZHAR HIRSCHFELD, GIDEON FOERSTER, and FANNY VITTO, "Tiberias," in EPHRAIM STERN, AYELET LEWINSON-GILBOA, and JOSEPH AVIRAM, *The New Encyclopedia of Archaeological Excavations in the Holy Land*, 4 vol. (1993; originally published in Hebrew, 1992), pp. 1464–73, are good, short introductions to the cities that take into account the ongoing archaeological work at the sites. REBECCA MARTIN NAGY *et al.* (eds.), *Sepphoris in Galilee: Crosscurrents of Culture* (1996), is a well-illustrated, general introduction to the history and culture of the city.

(E.P.Sa.)

Joan of Arc

Joan of Arc (French Jeanne d'Arc) was a peasant girl who, believing that she was acting under divine guidance, led the French army in a momentous victory at Orléans that repulsed an English attempt to conquer France during the Hundred Years' War. Captured a year afterward, Joan was burned by the English and their French collaborators as a heretic. She became the greatest national heroine of her compatriots. Her achievement was a decisive factor in the awakening of French national consciousness. She was canonized in 1920.

Giraudon—Art Resource/EB Inc.

St. Joan, equestrian miniature from the manuscript *La Vie des femmes célèbres* by Antoine Dufour, *c.* 1505. In the Musée Archéologique Thomas Dobrée, Nantes, Fr.

Joan was born *c.* 1412, the daughter of a tenant farmer, at Domrémy, on the borders of the duchies of Bar and Lorraine. In her mission of expelling the English and their Burgundian allies from the Valois kingdom of France, she felt herself to be guided by the "voices" of St. Michael, St. Catherine, and St. Margaret. She possessed many attributes characteristic of the female visionaries who were a noted feature of her time. These qualities included extreme personal piety, a claim to direct communication with the saints, and a consequent reliance upon individual experience of God's presence beyond the ministrations of the priesthood and the confines of the institutional church. But to these were added remarkable mental and physical courage, as well as a robust common sense. Known as La Pucelle, or the Maid of Orléans, Joan became in the following centuries a focus of unity for the French people, especially at times of crisis.

JOAN'S MISSION

The crown of France at the time was in dispute between the dauphin Charles, son and heir of the Valois king Charles VI, and the Lancastrian English king Henry VI. Henry's armies were in alliance with those of Philip the Good, duke of Burgundy (whose father, John the Fearless, had been assassinated in 1419 by partisans of the Dauphin), and were occupying much of the northern part of the kingdom. The apparent hopelessness of the Dauphin's cause at the end of 1427 was increased by the fact that, five years after his father's death, he still had not been crowned. Reims, the traditional place for the investiture of French kings, was well within the territory held by his enemies. As long as the Dauphin remained unconsecrated, the rightfulness of his claim to be king of France was open to challenge.

Joan's village of Domrémy was on the frontier between the France of the Anglo-Burgundians and that of the Dauphin. The villagers had already had to abandon their homes before Burgundian threats. Led by her voices, Joan traveled in May 1428 from Domrémy to Vaucouleurs, the nearest stronghold still loyal to the Dauphin, where she asked the captain of the garrison, Robert de Baudricourt, for permission to join the Dauphin. He did not take the 16-year-old girl and her visions seriously, and she returned home. Joan went to Vaucouleurs again in January 1429. This time her quiet firmness and piety gained her the respect of the people; and the captain, persuaded that she was neither a witch nor feebleminded, allowed her to go to the Dauphin at Chinon. She left Vaucouleurs about February 13, dressed in men's clothes and accompanied by six men-at-arms. Crossing territory held by the enemy, and traveling for 11 days, she reached Chinon.

Meeting with the Dauphin

Joan went at once to the castle occupied by the dauphin Charles. He was uncertain whether to receive her, and his counselors gave him conflicting advice; but two days later he granted her an audience. Charles had hidden himself among his courtiers, but Joan made straight for him and told him that she wished to go to battle against the English and that she would have him crowned at Reims. On the Dauphin's orders she was immediately interrogated by ecclesiastical authorities in the presence of Jean, duc d'Alençon, a relative of Charles, who showed himself well-disposed toward her. For three weeks she was further questioned at Poitiers by eminent theologians who were allied to the Dauphin's cause. These examinations, the record of which has not survived, were occasioned by the ever-present fear of heresy following the end of the Great Schism in 1417. Joan told the ecclesiastics that it was not at Poitiers but at Orléans that she would give proof of her mission; and forthwith, on March 22, she dictated letters of defiance to the English. In their report the churchmen suggested that in view of the desperate situation of Orléans, which had been under English siege for months, the Dauphin would be well-advised to make use of her.

Joan returned to Chinon. At Tours, during April, the Dauphin provided her with a military household of several men; Jean d'Aulon became her squire, and she was joined by her brothers Jean and Pierre. She had her standard painted with an image of Christ in Judgment and a banner made bearing the name of Jesus. When the question of a sword was brought up, she declared that it would be found in the church of Sainte-Catherine-de-Fierbois, and one was in fact discovered there.

Troops numbering several hundred men were mustered at Blois, and on April 27 they set out for Orléans. The city, besieged since Oct. 12, 1428, was almost totally surrounded by a ring of English strongholds. When Joan and one of the French commanders, La Hire, entered with supplies on April 29, she was told that action must be deferred until further reinforcements could be brought in.

Relief of Orléans

On the evening of May 4, when Joan was resting, she suddenly sprang up, apparently inspired, and announced that she must go and attack the English. Having herself armed, she hurried out to the east of the city toward an English fort where, indeed, an engagement of which she had not been told was taking place. Her arrival roused the French, and they took the fort. The next day Joan addressed another of her letters of defiance to the English. On the morning of May 6 she crossed to the south bank of the river and advanced toward another fort; the English immediately evacuated it in order to defend a stronger position nearby, but Joan and La Hire attacked them there and took it by storm. Very early on May 7 the French advanced against the fort of Les Tourelles. Joan was wounded but quickly returned to the fight, and it was thanks in part to her example that the French commanders maintained the attack until the English capitulated. Next day the English were seen to be retreating, but, because it was a Sunday, Joan refused to allow any pursuit.

Joan left Orléans on May 9 and met Charles at Tours. She urged him to make haste to Reims to be crowned. Though he hesitated because some of his more prudent counselors were advising him to undertake the conquest of Normandy, Joan's importunity ultimately carried the day. It was decided, however, first to clear the English out of the other towns along the Loire River. Joan met her friend the Duc d'Alençon, who had been made lieutenant general of the French armies, and they moved off together, taking a town and an important bridge. They next attacked Beaugency, whereupon the English retreated into the castle. Then, notwithstanding the opposition of the Dauphin and Georges de La Trémoille, one of his favourites, and despite the reserve of Alençon, Joan received the Constable de Richemont, who was under suspicion at the French court. After making him swear fidelity, she accepted his help. Shortly thereafter the castle of Beaugency was surrendered.

Battle of Patay

The French and English armies came face to face at Patay on June 18, 1429. Joan promised success to the French, saying that Charles would win a greater victory that day than any he had won so far. The victory was indeed complete; the English army was routed and with it, finally, its reputation for invincibility.

Instead of pressing home their advantage by a bold attack upon Paris, Joan and the French commanders turned back to rejoin the Dauphin, who was staying with La Trémoille at Sully-sur-Loire. Again Joan urged upon Charles the need to go on swiftly to Reims. He vacillated, however; and as he meandered through the towns along the Loire, Joan accompanied him, arguing all the while in an attempt to vanquish his hesitancy and prevail over the counselors who advised delay. She was not unaware of the dangers and difficulties involved but declared them of no account. Finally she won Charles to her view.

From Gien, where the army began to assemble, the Dauphin sent out the customary letters of summons to the coronation. Joan wrote two letters: one of exhortation to the people of Tournai, always loyal to Charles, the other a challenge to Philip the Good, duke of Burgundy. She and the Dauphin set out on the march to Reims on June 29. Before arriving at Troyes, Joan wrote to the inhabitants, promising them pardon if they would submit. They countered by sending a friar, the popular preacher Brother Richard, to take stock of her; but though he returned full of enthusiasm for the Maid and her mission, the townsfolk decided after all to remain loyal to the Anglo-Burgundian regime. The Dauphin held a council, and Joan proposed that the town be attacked. The next morning she began the assault, and the citizens at once asked for terms. The royal army then marched on to Châlons. Despite an earlier decision to resist, the Count-Bishop handed the keys of the town to Charles. On July 16 the royal army reached Reims, which opened its gates. The coronation took place on July 17, 1429. Joan was present at the consecration, standing with her banner not far from the altar. After the ceremony she knelt before Charles, calling him her king for the first time. That same day she wrote to the Duke of Burgundy, adjuring him to make peace with the King and to withdraw his garrisons from the royal fortresses.

Coronation of the Dauphin

Charles VII left Reims on July 20, and for a month the army paraded through Champagne and the Île-de-France. On August 2 the King decided on a retreat from Provins to the Loire, a move that implied abandoning any plan to attack Paris. The loyal towns that would thus have been left to the enemy's mercy expressed some alarm. Joan, who was opposed to Charles's decision, wrote to reassure the citizens of Reims on August 5, saying that the Duke of Burgundy, then in possession of Paris, had made a fortnight's truce, after which it was hoped that he would yield Paris to the King. In fact, on August 6, English troops prevented the royal army from crossing the Seine at Bray, much to the delight of Joan and the commanders, who hoped that Charles would attack Paris. Everywhere acclaimed, Joan was now, according to a 15th-century chronicler, the idol of the French. She herself felt that the purpose of her mission had been achieved.

Near Senlis, on August 14, the French and English armies again confronted each other. This time only skirmishes took place, neither side daring to start a battle, though Joan carried her standard up to the enemy's earthworks and openly challenged them. Meanwhile Compiègne, Beauvais, Senlis, and other towns north of Paris surrendered to the King. Soon afterward, on August 28, a four months' truce for all the territory north of the Seine was concluded with the Burgundians.

Attack on Paris

Joan, however, was becoming more and more impatient; she thought it essential to take Paris. She and Alençon were at Saint-Denis on the northern outskirts of Paris on August 26, and the Parisians began to organize their defenses. Charles arrived on September 7, and an attack was launched on September 8, directed between the gates of Saint-Honoré and Saint-Denis. The Parisians could be in no doubt of Joan's presence among the besiegers; she stood forward on the earthworks, calling on them to surrender their city to the King of France. Wounded, she continued to encourage the soldiers until she had to abandon the attack. Though the next day she and Alençon sought to renew the assault, they were ordered by Charles's council to retreat.

Charles VII retired to the Loire, Joan following him. At Gien, which they reached on September 22, the army was disbanded. Alençon and the other captains went home; only Joan remained with the King. Later, when Alençon was planning a campaign in Normandy, he asked the King to let Joan rejoin him, but La Trémoille and other courtiers dissuaded him. Joan went with the King to Bourges, where many years later she was to be remembered for her goodness and her generosity to the poor. In October she was sent against Saint-Pierre-le-Moûtier; through her courageous assault, with only a few men, the town was taken. Joan's army then laid siege to La Charité-sur-Loire; short of munitions, they appealed to neighbouring towns for help. The supplies arrived too late, and after a month they had to withdraw.

Compiègne

Joan then rejoined the King, who was spending the winter in towns along the Loire. Late in December 1429 Charles issued letters patent ennobling Joan, her parents, and her brothers. Early in 1430 the Duke of Burgundy began to threaten Brie and Champagne. The inhabitants of Reims became alarmed, and Joan wrote in March to assure them of the King's concern and to promise that she would come to their defense. When the Duke moved up to attack Compiègne, the townsfolk determined to resist, and in late March or early April Joan left the King and set out to their aid, accompanied only by her brother Pierre, her squire Jean d'Aulon, and a small troop of men-at-arms. She arrived at Melun in the middle of April, and it was no doubt her presence that prompted the citizens there to declare themselves for Charles VII.

Joan was at Compiègne by May 14, 1430. There she found Renaud de Chartres, archbishop of Reims, and Louis I de Bourbon, comte de Vendôme, a relative of the King. With them she went on to Soissons, where the townspeople refused them entry. Renaud and Vendôme therefore decided to return south of the Marne and Seine rivers; but Joan refused to accompany them, preferring to return to her "good friends" in Compiègne.

CAPTURE, TRIAL, AND EXECUTION

On her way back Joan heard that John of Luxembourg, the captain of a Burgundian company, had laid siege to Compiègne. Hurrying on, she entered Compiègne under cover of darkness. The next afternoon, May 23, she led a sortie and twice repelled the Burgundians but was eventually outflanked by English reinforcements and compelled to retreat. Remaining until the last to protect the rear guard while they crossed the Oise River, she was unhorsed and could not remount. She gave herself up and, with her brother Pierre and Jean d'Aulon, was taken to Margny, where the Duke of Burgundy came to see her. In telling the people of Reims of Joan's capture, Renaud de Chartres accused her of rejecting all counsel and acting willfully.

Charles, who was working toward a truce with the Duke of Burgundy, made no attempts to save her.

Imprisonment

John of Luxembourg sent Joan and Jean d'Aulon to his castle in Vermandois. When she tried to escape in order to return to Compiègne, he sent her to one of his more distant castles. There, though she was treated kindly, she became more and more distressed at the predicament of Compiègne. Her desire to escape became so great that she jumped from the top of a tower, falling unconscious into the moat. She was not seriously hurt, and when she had recovered, she was taken to Arras, a town adhering to the Duke of Burgundy.

News of her capture had reached Paris on May 25. Next day the University of Paris, which had taken the English side, requested the Duke of Burgundy to turn her over for judgment either to the chief inquisitor or to the bishop of Beauvais, Pierre Cauchon, in whose diocese she had been seized. The university wrote also, to the same effect, to John of Luxembourg; and on July 14 the Bishop of Beauvais presented himself before the Duke of Burgundy asking, on his own behalf and in the name of the English king, that the Maid be handed over in return for a payment of 10,000 francs. The Duke passed on the demand to John of Luxembourg, and by Jan. 3, 1431, she was in the Bishop's hands. The trial was fixed to take place at Rouen. Joan was moved to a tower in the castle of Bouvreuil, which was occupied by the Earl of Warwick, the English commander at Rouen. Though her offenses against the Lancastrian monarchy were common knowledge, Joan was brought to trial before a church court because the University of Paris, as arbiter in matters concerning the faith, insisted that she be tried as a heretic. Her beliefs were not strictly orthodox, according to the criteria for orthodoxy laid down by many theologians of the period. She was no friend of the church militant on Earth (which perceived itself as in spiritual combat with the forces of evil), and she threatened its hierarchy through her claim that she communicated directly with God by means of visions or voices. Further, her trial might serve to discredit Charles VII by demonstrating that he owed his coronation to a witch, or at least a heretic. Her two judges were to be Cauchon, bishop of Beauvais, and Jean Lemaître, the vice-inquisitor of France.

The trial. Beginning Jan. 13, 1431, statements taken in Lorraine and elsewhere were read before the Bishop and his assessors; they were to provide the framework for Joan's interrogation. Summoned to appear before her judges on February 21, Joan asked for permission to attend mass beforehand, but it was refused on account of the gravity of the crimes with which she was charged, including attempted suicide in having jumped into the moat. She was ordered to swear to tell the truth and did so swear, but she always refused to reveal the things she had said to Charles. Cauchon forbade her to leave her prison, but Joan insisted that she was morally free to attempt escape. Guards were then assigned to remain always inside the cell with her, and she was chained to a wooden block and sometimes put in irons. Between February 21 and March 24 she was interrogated nearly a dozen times. On every occasion she was required to swear anew to tell the truth, but she always made it clear that she would not necessarily divulge everything to her judges since, although nearly all of them were Frenchmen, they were enemies of King Charles. The report of this preliminary questioning was read to her on March 24, and apart from two points she admitted its accuracy.

Interrogation

When the trial proper began a day or so later, it took two days for Joan to answer the 70 charges that had been drawn up against her. These were based mainly on the contention that her whole attitude and behaviour showed blasphemous presumption: in particular, that she claimed for her pronouncements the authority of divine revelation; prophesied the future; endorsed her letters with the names of Jesus and Mary, thereby identifying herself with the novel and suspect cult of the Name of Jesus; professed to be assured of salvation; and wore men's clothing. Perhaps the most serious charge was of preferring what she believed to be the direct commands of God to those of the church.

On March 31 she was questioned again on several points about which she had been evasive, notably on the question of her submission to the church. In her position, obedience to the court that was trying her was inevitably made a test of such submission. She did her best to avoid this trap, saying she knew well that the church militant could not err, but it was to God and to her saints that she held herself answerable for her words and actions. The trial continued, and the 70 charges were reduced to 12, which were sent for consideration to many eminent theologians in both Rouen and Paris.

Meanwhile, Joan fell sick in prison and was attended by two doctors. She received a visit on April 18 from Cauchon and his assistants, who exhorted her to submit to the church. Joan, who was seriously ill and obviously thought she was dying, begged to be allowed to go to confession and receive Holy Communion and to be buried in consecrated ground. But they continued to badger her, receiving only her constant response "I am relying on our Lord," "I hold to what I have already said." They became more insistent on May 9, threatening her with torture if she did not clarify certain points. She answered that even if they tortured her to death she would not reply differently, adding that in any case she would afterward maintain that any statement she might make had been extorted from her by force. In face of this commonsense fortitude her interrogators, by a majority of 10 to three, decided on May 12 that torture would be useless. Joan was informed on May 23 of the decision of the University of Paris that if she persisted in her errors she would be turned over to the secular authorities; only they, and not the church, could carry out the death sentence of a condemned heretic.

Abjuration, relapse, and execution. Apparently nothing further could be done. Joan was taken out of prison for the first time in four months on May 24 and conducted to the cemetery of the church of Saint-Ouen, where her sentence was to be read out. First she was made to listen to a sermon by one of the theologians in which he violently attacked Charles VII, provoking Joan to interrupt him because she thought he had no right to attack the King, a "good Christian," and should confine his strictures to her. After the sermon was ended, she asked that all the evidence on her words and deeds be sent to Rome. But her judges ignored her appeal to the Pope, to whom, under God, she would be answerable, and began to read out the sentence abandoning her to the secular power. Hearing this dreadful pronouncement, Joan quailed and declared she would do all that the church required of her. She was presented with a form of abjuration, which must already have been prepared. She hesitated in signing it, eventually doing so on condition that it was "pleasing to our Lord." She was then condemned to perpetual imprisonment or, as some maintain, to incarceration in a place habitually used as a prison. In any case, the judges required her to return to her former prison.

The vice-inquisitor had ordered Joan to put on women's clothes, and she obeyed. But two or three days later, when the judges and others visited her and found her again in male attire, she said she had made the change of her own free will, preferring men's clothes. They then pressed other questions, to which she answered that the voices of St. Catherine and St. Margaret had censured her "treason" in making an abjuration. These admissions were taken to signify relapse, and on May 29 the judges and 39 assessors agreed unanimously that she must be handed over to the secular officials.

Revocation of abjuration

The next morning, May 30, 1431, Joan received from Cauchon permission, unprecedented for a relapsed heretic, to make her confession and receive Communion. Accompanied by two Dominicans, she was then led to the Place du Vieux-Marché. There she endured one more sermon, and the sentence abandoning her to the secular arm—that is, to the English and their French collaborators—was read out in the presence of her judges and a great crowd. The executioner seized her, led her to the stake, and lit the pyre. A Dominican consoled Joan, who asked him to hold high a crucifix for her to see and to shout out the assurances of salvation so loudly that she should hear him above the roar of the flames. To the last she maintained that her voices were sent of God and had not deceived her. According to the rehabilitation proceedings of 1456, few witnesses of her death seem to have doubted her salvation, and they agreed that she died a faithful Christian. A few days later the English king and the University of Paris formally published the news of Joan's execution.

Rehabilitation

Almost 20 years afterward, on his entry into Rouen in 1450, Charles VII ordered an inquiry into the trial. Two years later the cardinal legate Guillaume d'Estouteville made a much more thorough investigation. Finally, on the order of Pope Calixtus III following a petition from the d'Arc family, proceedings were instituted in 1455–56 that revoked and annulled the sentence of 1431. Joan was canonized by Pope Benedict XV on May 16, 1920; her feast day is May 30. The French parliament, on June 24, 1920, decreed a yearly national festival in her honour; this is held the second Sunday in May.

CHARACTER AND IMPORTANCE

Joan of Arc's place in history is assured. Perhaps her contribution to the history of human courage is greater than her significance in the political and military history of France. She was victimized as much by a French civil conflict as by a war with a foreign power. The relief of Orléans was undoubtedly a notable victory, which secured the loyalty of certain regions of northern France to the régime of Charles VII. But the Hundred Years' War continued for a further 22 years after her death, and it was the defection of Philip the Good of Burgundy from his alliance with the Lancastrians in 1435 that provided the foundation upon which the recovery of Valois France was to be based. The nature of Joan's mission, moreover, is a source of controversy among historians, theologians, and psychologists. Innumerable points about her campaigns and about the motives and actions of her supporters and enemies are subject to dispute: for instance, the number and dates of her visits to Vaucouleurs, Chinon, and Poitiers; how she was able to win the confidence of the Dauphin at their first meeting at Chinon; whether Charles's perambulations after his coronation at Reims represented triumphant progress or scandalous indecision; what her judges meant by "perpetual imprisonment"; whether, after her recantation, Joan resumed men's clothes of her own free will and at the bidding of her voices or, as one later story has it, because they were forced upon her by her English jailers.

Assessment

Later generations have tended to distort the significance of Joan's mission according to their own political and religious viewpoints rather than seeking to set it in the troubled context of her time. The effects of the Great Schism within the Western Church (1378–1417) and the decline of papal authority during the Conciliar Movement (1409–49) made it difficult for persons to seek independent arbitration and judgment in cases relating to the faith. The verdicts of the Inquisition were liable to be coloured by political and other influences; and Joan was not the only victim of an essentially unjust procedure, which allowed the accused no counsel for the defense and which sanctioned interrogation under duress. Her place among the saints is secured, not perhaps by the somewhat dubious miracles attributed to her, but by the heroic fortitude with which she endured the ordeal of her trial and, except for one lapse toward its end, by her profound conviction of the justice of her cause, sustained by faith in the divine origin of her voices. In many ways a victim of internal strife within France, condemned by judges and assessors who were almost entirely northern French in origin, she has become a symbol of national consciousness with whom all French people, of whatever creed or party, can identify.

BIBLIOGRAPHY. Thereis an extremely large literature on Joan of Arc but no truly definitive biography. Important works include: JULES QUICHERAT (ed.), *Procès de condamnation et de réhabilitation de Jeanne d'Arc*, 5 vol. (1841–49, reprinted 1965); *Le Procès de condamnation de Jeanne d'Arc*, 2 vol., trans. and annotated by PIERRE TISSET (1960–70); PAUL DONCOEUR (ed.), *La Minute française des interrogatoires de Jeanne la Pucelle* (1952); *La Réhabilitation de Jeanne la Pucelle*, 3 vol., trans. and annotated by PAUL DONCOEUR and YVONNE LANHERS (1956–61), containing the recorded testimony from the trial of more than 100 witnesses who had known Joan of Arc; and PIERRE LANÉRY D'ARC, *Les Mémoires et consultations en faveur de Jeanne d'Arc* (1889), dealing with the memoirs of jurists and theologians, intended to facilitate her rehabilitation. In English, DANIEL S. RANKIN and CLAIRE QUINTAL (eds. and trans.), *The First Biography of Joan of Arc with the Chronicle Record of a Contemporary Account* (1964); and RÉGINE PERNOUD, *Joan of Arc by Herself and Her Witnesses* (1964, reissued 1982; originally published in French, 1962), and *The Retrial of Joan of Arc: The Evidence at the Trial for Her Rehabilitation* (1955; originally published in French, 1953), are particularly recommended. Other works include GEORGES DUBY and ANDRÉE DUBY (eds.), *Les Procès de Jeanne d'Arc* (1973); JOHN H. SMITH, *Joan of Arc* (1973), which explores her military activities; HENRI GUILLEMIN, *Joan, Maid of Orléans* (1973; originally published in French, 1970); PHILIPPE WOLFF, "Le Théologien Pierre Cauchon, de sinistre mémoire," in *Économies et sociétés au Moyen Âge: Mélanges offerts à Edouard Perroy*, pp. 553–570 (1973); EDWARD LUCIE-SMITH, *Joan of Arc* (1976); WALTER S. SCOTT, *Jeanne d'Arc* (1974); MARINA WARNER, *Joan of Arc: The Image of Female Heroism* (1981, reprinted 1982), a psychohistorical approach with an important survey of the posthumous history of Joan's legend; FRANCES GIES, *Joan of Arc: The Legend and the Reality* (1981), an examination of her life and the literature about her; and COLLOQUE D'HISTOIRE MÉDIÉVALE. ORLÉANS, FRANCE, 1979, *Jeanne d'Arc: une époque, un rayonnement* (1982), an informative collection of scholarly papers. For the political background to Joan's career see M.G.A. VALE, *Charles VII* (1974); C.T. ALLMAND, *Lancastrian Normandy, 1415–1450* (1983); and ROGER G. LITTLE, *The Parlement of Poitiers: War, Government and Politics in France, 1418–1436* (1984).

(Y.L./M.G.A.V.)

Johannesburg

Johannesburg is South Africa's largest metropolis as well as its industrial and financial engine. One of the youngest of the world's major cities, it was founded in 1886, following the discovery of gold. The city was initially part of the Transvaal, an independent Afrikaner, or Boer, republic that later became one of the four provinces of South Africa. Today the city belongs to Gauteng (a Sotho word meaning "Place of Gold"), the most populous of the nine provinces in the Republic of South Africa.

The geography of Johannesburg reflects nearly a century of racially driven social engineering that reached a climax under apartheid (literally "apartness"), the system of racial segregation that obtained in South Africa from 1948 to 1994. The result is a city of extraordinary contrasts, of glass and steel skyscrapers and fetid shantytowns, of internationally recognized universities and widespread illiteracy, of glittering abundance and desperate poverty.

The article is divided into the following sections:

Physical and human geography

THE LANDSCAPE

The city site. Johannesburg is situated on the Highveld, the broad, grassy plateau that sweeps across the South African interior. The city bestrides the Witwatersrand, or Rand, a string of low, rocky ridges that constitutes the watershed between the drainages into the Indian and Atlantic oceans. The city's elevation ranges from 5,700 to 5,930 feet (1,740 to 1,810 metres).

Aside from a few small streams and artificial lakes, Johannesburg lacks water; indeed, it has the distinction of being the world's largest city not on a navigable body of water. The city owes its location to the presence of an even more precious resource: gold. The city grew on the edge of the Witwatersrand Main Reef, a subterranean stratum of gold-bearing quartz-silica conglomerate that arcs for hundreds of miles beneath the Highveld. Most of the gold mines in the city ceased operation in the 1970s, but in its day the Witwatersrand gold industry accounted for more than 40 percent of the world's annual gold production. Remnants of the industry—rusting headgear, towering yellow-white mine dumps, copses of dusty Australian bluegum trees imported for underground timbering—still litter the landscape.

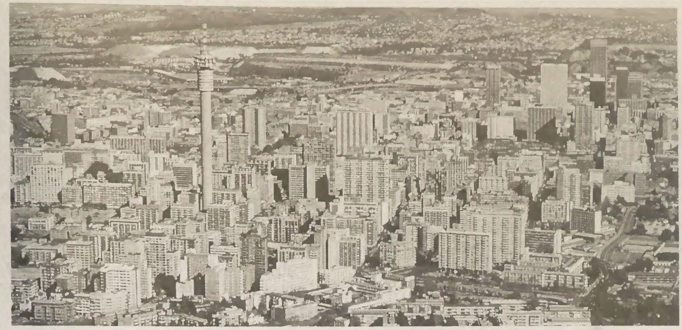
Climate. Johannesburg has a temperate climate. Summertime temperatures average about 75° F (24° C); winter temperatures average about 55° F (13° C) and only occasionally dip below freezing. Rainfall averages about 28 inches (700 millimetres) per annum, but the total varies considerably from year to year. Droughts are common. What rain the city receives falls almost exclusively in the summer months, often in spectacular late-afternoon electrical storms. Air pollution poses a significant problem, especially in the winter months, when thermal inversions impede the westward flow of air from the Indian Ocean. Pollution is most severe in the African townships on the city's periphery, where many residents still rely on coal for fuel.

The city layout. Central Johannesburg, the commercial and financial heart of South Africa, is laid out in a rectangular grid pattern that is unchanged from the first city survey in 1886. Streets are narrow and cast into shadow by high-rise concrete blocks, creating an almost tunnellike effect. Architecturally, the city is a hodgepodge, reflecting decades of rapid growth and a singular indifference to historic preservation. The tents and clay huts of the original mining camp are gone, as are most of the ornate, gabled Victorian edifices that sprang up in the 1890s. (Markhams Building, on Pritchard Street, is a conspicuous exception.) The early 20th century brought a variety of architectural styles and movements. Monumental Beaux Arts structures such as the Supreme Court building and the Johannesburg Art Gallery bespoke the city's new status as an outpost of the British Empire, while massive, steel-reinforced concrete blocks such as Corner House, headquarters of one of South Africa's leading mining houses, reflected the growing importance of American architectural techniques and idioms. American influence was even more apparent in the 1930s "skyscraper" movement, most notably in the 1937 ESKOM Building, a 21-story Art Deco tower built to evoke the vigour of New York City. (The ESKOM Building was torn down in 1983, joining a distinguished line of vanished landmarks.) Whatever architectural distinction the city had was lost in the decades after World War II amidst a sea of nondescript high-rise blocks.

Greater Johannesburg, an area of more than 200 square miles, comprises more than 500 suburbs and townships. Under the terms of the 1950 Group Areas Act, the cornerstone of urban apartheid (see below), each was reserved for a single "race group." The act was repealed in 1991, but Johannesburg retains a high degree of racial segregation.

Africans can be found throughout the city, but the majority still live in "townships" on the urban periphery, essentially dormitory cities for blacks working in the city. Townships
Alexandra township, a 20-square-block enclave carved out of Johannesburg's white northern suburbs, houses a population of nearly half a million. At least three times that number live in Soweto (South-West Townships), a sprawling urban complex 10 miles southwest of the city. Johannesburg's small Coloured population (people of mixed race) clusters in townships west of the city, while the bulk of its Indian population lives in Lenasia, a special "Asiatic" township built in the 1950s to accommodate Indians forcibly removed from the city centre. The balance of the city is occupied by whites.

Accommodation varies in character and quality. Soweto is notorious for its endless rows of municipally built, two-room matchbox homes, yet it also has a few prosperous enclaves as well as teeming squatter camps, where tens of thousands live without water, electricity, or sanitation facilities. African migrant workers, long the backbone of South Africa's industrial labour force, are lodged in massive, single-sex hostels located close to the workplace or on the edge of black townships. White accommodation varies from suburb to suburb. In western suburbs such as



Central Johannesburg, with the J.G. Strijdom (or Hillbrow) Tower for telecommunications at centre left. Gold mine dumps are visible in the background.
© Joyce Photographics/Photo Researchers

Brixton and Melville, middle-class whites live in the modest tin-roofed bungalows and semidetached homes that once housed the city's white working class. Conditions are bleaker in neighbouring suburbs such as Cottesloe, Vrededorp, and Booysens Reserve, home to most of Johannesburg's white poor. More affluent whites live in the north, in leafy, established communities such as Houghton and Parktown, once the residence of South Africa's mining magnates, or in any of a dozen newer suburbs. Northern suburban homes typically include large, flowering gardens and swimming pools. Most are surrounded by high fences.

THE PEOPLE

About 70 percent of Johannesburg's citizens are African; about 25 percent are white; most of the remainder are Coloured or Indian. Such figures, however, scarcely do justice to the city's polyglot population. At least a dozen different languages are in widespread, daily use in Johannesburg. The majority of the white population is of English and Afrikaans descent, but the city also includes substantial Portuguese, Greek, Italian, Russian, Polish, and Lebanese communities. The African population includes representatives from every ethnic and linguistic group in southern Africa. All the world's major religions are represented, though the majority of the people, both white and black, are Christian.

Religions

THE ECONOMY

Commerce and industry. Johannesburg is a centre of mining, manufacturing, and finance. All the mining houses are headquartered in the city, as is the Chamber of Mines, which regulates the industry. Local factories in Johannesburg and on the East Rand produce a great variety of goods ranging from textiles to specialty steels. A substantial engineering sector serves the needs of the mining industry. Virtually all the country's banks, insurance companies, and building societies have their head offices in the city. The Johannesburg Stock Exchange, founded in 1887 to raise capital for deep-level mining, lists more than 600 companies.

While Pretoria, the South African capital, is only 40 miles to the north, most state ministries have offices in Johannesburg. Many foreign countries retain consular facilities, largely to service the needs of overseas firms, hundreds of which operate in the city.

Transportation. Johannesburg is a hub for local, national, and international travel. Railroads and multilane freeways crisscross the metropolitan area, carrying hundreds of thousands of daily commuters to and from outlying suburbs and townships. A municipal bus system operates within the city, and a separate, private bus company operating under a state monopoly connects the city centre with Soweto and Alexandra. The inadequacies of the latter have fostered a burgeoning local taxi industry. Johannesburg International Airport, 14 miles northeast of the city centre, offers regularly scheduled service between Johannesburg and most cities in southern Africa, as well as direct flights to Europe, North and South America, Australia, and Asia.

ADMINISTRATION AND SOCIAL CONDITIONS

Government. Under the new constitution, which went into effect in 1994, South Africa possesses a federal system of government, with authority divided between national, provincial, and local levels of government. Local authority for Johannesburg rests with the Greater Johannesburg Metropolitan Board, which includes representatives from all across the metropolitan area. In extending the municipal borders to include previously disfranchised black townships such as Soweto and Alexandra, political leaders hope to facilitate some equalization of revenues and services between white and black areas.

Public services. Local bus service, fire fighting, and sanitation remain the province of the municipality, but other responsibilities—such as the provision of housing in African townships—have been assumed by the Gauteng regional legislature. Electricity is provided through the Electricity Supply Commission (ESKOM), a national parastatal institution. Water is supplied by the Rand Water Board. Municipal police oversee traffic control; other policing is provided by the South African Police Services. The South African Broadcasting Corporation is headquartered in Auckland Park, west of the city centre.

Health. Apartheid left perhaps its deepest imprint on public health. In black townships rates of child mortality are significantly higher and life spans far shorter than in white neighbourhoods. Tuberculosis, virtually eradicated among whites, remains endemic in townships and migrant hostels. Addressing such historical inequities is one of the chief challenges of the national government.

Education. Primary and secondary schools range widely in character. Racial segregation, abolished in law, remains common in practice. Facilities of higher education abound. The University of the Witwatersrand is South Africa's premier university. Founded in 1896 as the South African School of Mines, "Wits" today confers degrees in commerce, arts, sciences, architecture, law, education, medicine, and dentistry. Rand Afrikaans University (RAU), an Afrikaans-medium university, opened in 1968. Vista University opened a branch in Soweto in 1982. Teacher-training courses are given at five different colleges of education, each originally earmarked for a specific racial or linguistic group. Tertiary technical education is available at Technikon Witwatersrand.

Johannesburg and surrounding area.

CULTURAL LIFE

Johannesburg boasts a rich, if strangely schizophrenic, cultural life. Patrons of the arts can take in a performance of the National Symphony Orchestra and then retire to one of the city's thriving jazz clubs to hear internationally acclaimed local performers, many of whom have returned to Johannesburg after long years in exile. The Johannesburg Art Gallery, established in the early years of the 20th century with donations from mining magnates, features Africa's finest collection of European Impressionists, while most of the city's dozen private galleries increasingly highlight the work of African artists. Theatre flourishes. While the 1,100-seat Civic Theatre stages European operas and American musicals, many smaller companies nurture the talents of local actors and playwrights, white and black. A visit to Johannesburg is incomplete without a trip to the world-renowned Market Theatre, a multitheatre complex housed in the city's old produce market.

The city has many museums and libraries. Johannesburg Public Library, first established in 1889, is the centre of an extensive network of branch libraries. Local museums specialize in geology, Africana, military history, archaeology, transport, banking, costume, and Judaica. Visitors interested in a taste of old Johannesburg can visit Gold Reef City, an amusement park located a few miles south of the city on the site of a defunct gold mine. Those interested in a less sanitized version of the city's past can visit the Museum Africa or the Bensusan Museum of Photography, both at the old market.

Gold Reef City

Blessed as they are with a warm, sunny climate, Johannesburgers spend a considerable amount of time outdoors. The northern suburbs feature broad swaths of open space for bird-watching and picnicking. Weekend cookouts—*braaivleis,* in local parlance—are a summertime ritual, especially among Afrikaners. Like other South Africans, Johannesburg residents tend to be avid sportsmen. In the years since South Africa's return from political isolation, the city has hosted international competitions in rugby, soccer, cricket, golf, and tennis.

History

THE EARLY PERIOD, 1853–1930

Boomtown. Johannesburg's early history is the story of gold. In 1853 Pieter Jacob Marais, a South African prospector, recovered alluvial gold from the Jukskei River, north of what would become Johannesburg. The years that followed brought several modest strikes, but the Witwatersrand Main Reef eluded searchers until 1886, when George Harrison, an Australian prospector, chanced upon an outcropping on a farm called Langlaagte. Ironically, Harrison failed to appreciate the significance of his find: he sold his claim for £10 and embarked for the goldfields of the eastern Transvaal region.

Founding of Johannesburg

Others were more farsighted. By mid-1886 an army of diggers had descended on the Witwatersrand, hacking away with picks and shovels along a line that soon stretched 40 miles west to east. In response to this influx, the government of the Transvaal, the small Boer republic under whose jurisdiction the Witwatersrand fell, dispatched two men, Vice President Christiaan Johannes Joubert and Deputy Surveyor-General Johann Rissik, to inspect the goldfields and identify a suitable city site. The new city was called Johannesburg, apparently in their honour.

As the scale of the gold deposits became apparent, Johannesburg became the 19th century's last great boomtown. Fortune hunters from as far afield as Australia and California joined skilled Cornish and Welsh miners, who brought to South Africa a strong trade-union tradition. Destitute Afrikaners, driven from their rural homes by debt and drought, clustered in slums such as Brickfields and Vrededorp. Africans from every corner of the southern African subcontinent migrated to the city, often in large ethnic cohorts, adding a dozen more voices to the cultural and linguistic babel. Most Africans worked on the mines, completing six- and nine-month contracts before returning to their rural homes. Others settled permanently in the swelling city, carving out niches as rickshaw drivers, domestic workers, and washermen. By 1896 Johannesburg had become a city of 100,000 people.

Conceived in avarice, the young city nurtured every species of vice. Banks and boardinghouses jostled for space with more than 500 saloons. Criminal syndicates with roots in New York City and London found fertile soil in Johannesburg. The predominantly male population provided a robust market for prostitution. "Ancient Ninevah and Babylon have been revived," a visiting journalist wrote in 1913. "Johannesburg is their twentieth century prototype. It is a city of unbridled squalor and unfathomable squander."

Consolidation of the gold industry. The gold deposits of the Main Reef, for all their uncanny dependability, were also extremely low-grade. Tons of the pebbly conglomerate had to be mined, crushed, amalgamated with mercury (later cyanide), and retorted in order to produce even an ounce or two of gold. This fact, combined with gold's internationally fixed price, produced a perennial problem of profitability, which increased exponentially as the reef dipped away to the south to depths of hundreds, and ultimately thousands, of feet. (South African gold mines would eventually reach depths of over two miles, making them far and away the deepest mines in the world.)

All these factors promoted a rapid consolidation of the industry. By the mid-1890s control of the entire Witwatersrand gold industry rested in the hands of a half-dozen massive mining houses, each of which commanded thousands of workers and millions of dollars in capital, most of it raised from investors in Europe and the United States. Control of these companies lay with a small number of so-called "Randlords," men such as Alfred Beit, Barney Barnato, and J.B. Robinson, who had made their fortunes on the Kimberley diamond fields and well understood the exigencies of large-scale industrial mining. Working under the auspices of the newly formed Chamber of Mines, the Randlords strove to establish the profitability of their industry by rationalizing production and relentlessly squeezing down costs, especially the cost of labour.

The national and international context. As the Randlords' power waxed, so did their frustration with the Transvaal government, which they regarded as too corrupt and inefficient to meet the needs of a modern industrial economy. Boer officials extracted hefty bribes and handed out valuable concessions on supplies to political allies. Worse, the Transvaal government seemed unable to enact or enforce the kind of discriminatory taxes and rigorous master-servant laws that the Randlords regarded as essential to their campaign to reduce African labour costs. As one exasperated industry expert put it, Boers lacked the ability "to understand capitalism, industrialisation and progress."

The South African War

Mineowners' frustrations were stoked by British officials, many of whom were eager to see the goldfields brought within the orbit of the British Empire. In 1895, British officials tacitly endorsed the Jameson Raid, a coup attempt against the Transvaal government conceived by the mining magnate Cecil John Rhodes. When that failed, they seized on the plight of the "uitlanders"—the foreign, mostly British, miners in Johannesburg, who were denied the right to vote under Transvaal law. In September 1899 the British government delivered an ultimatum to the Boers demanding the immediate enfranchisement of all (white) uitlanders. In October 1899 the South African War (also known as the Boer War) began. When the fighting ceased two and a half years later, the Transvaal and its sister republic, the Orange Free State, were colonies of the British Empire.

British troops entered Johannesburg unopposed in June 1900. The mines, left undamaged by retreating Boers, were back in operation by the end of 1901. As mineowners had hoped, the Transvaal's new imperial overlords were sensitive to the industry's needs, rescinding Boer tariffs and concessions and enacting onerous new taxes and a pass law explicitly designed to force Africans to accept employment at whatever wages whites were willing to pay. When these devices failed to produce a sufficient pool of cheap labour, imperial officials cooperated with the Chamber of Mines in the temporary importation of more than 60,000 Chinese indentured labourers. By the inauguration of the Union of South Africa in 1910, the gold industry rested on a firm financial footing.

The local level. *Racist enactments.* The first decades of the 20th century were a time of extensive social engineering as municipal authorities, influenced by new currents in eugenics and city planning, attacked what they took to be the sources of urban disorder. In 1904, Africans living near the city centre were forcibly relocated to Klipspruit, 10 miles southwest of town. As had happened in earlier removals in Cape Town and Port Elizabeth, the move was preceded by a plague scare and accomplished in the name of "sanitation," though it is difficult to see how the interests of sanitation were served at Klipspruit, a municipal sewerage farm. Similarly strained justifications were used in ensuing decades to relocate not only Africans but Indians, Coloureds, and even poor whites. The process was facilitated by the ideology of segregation, which emerged in the first quarter of the 20th century as a kind of panacea for South Africa's "race problem." The 1923 Natives (Urban Areas) Act, for example, defined urban Africans as "temporary sojourners," welcome only insofar as they ministered "to the wants of the white population." While Johannesburg never availed itself of the full range of powers the law afforded, it took advantage of the act to relocate thousands of Africans from slums and backyards in the city to black townships such as Alexandra or Western Areas or to the new development at Orlando East, the first piece in the vast urban jigsaw that would become Soweto. Pass and liquor raids became regular features of township life, as police strove to root out those who were "idle," "disorderly," or simply superfluous to the white economy.

Black protests. Blacks did not simply accede to racist enactments. Africans organized innumerable petitions and deputations in the early years of the 20th century, protesting what they regarded as a betrayal of the British tradition of equality before the law. Protest exploded into outright resistance in the last years of World War I. Ground between low wages and murderous wartime inflation, African railway and municipal workers in Johannesburg staged bitter strikes. An emboldened Transvaal Native Congress, the local affiliate of the South African Native National Congress (renamed African National Congress [ANC] in 1923), launched a major antipass campaign, leading to several violent clashes with police. In 1920, 70,000 black mineworkers struck for better wages and working conditions. Eventually the army was called out to march the strikers underground at bayonet point.

Challenge by white workers. As significant as such resistance was, the chief challenge to South Africa's political order in the early decades of the 20th century came from Johannesburg's white workers. By custom, whites performed all skilled labour in the mines (a practice enshrined in the "job colour bar" provisions of the 1911 Mines and Works Act), commanding in return wages 5 to 10 times higher than those earned by Africans. Mine owners, ever on the lookout for ways to reduce costs, challenged white workers on three separate occasions—in 1907, 1913, and 1922—encountering violent resistance each time. In the last case, a plan to reduce the proportion of whites in the mine labour force touched off a general strike, which quickly escalated into a full-scale armed insurrection. White workers, marching under banners that proclaimed "Workers of the World Fight and Unite for a White South Africa," seized control of the entire city, surrendering only after the arrival of 20,000 troops and a sustained air and artillery bombardment.

The Rand Revolt

THE TRANSFORMATIONS OF THE 1930S AND '40S

In the 1930s and '40s Johannesburg was transformed by a massive in-migration of Africans from the countryside. Primarily a consequence of deteriorating conditions in rural reserves, African urbanization also reflected the increasing availability of employment in the Witwatersrand's rapidly growing manufacturing industries. Born of the exigencies of war and sustained through the 1920s by government tariffs, South African manufacturing exploded in the 1930s, especially in the boom years that followed the country's departure from the gold standard in 1933. By the early 1940s manufacturing had outstripped both mining and agriculture in terms of its contribution to the gross national product. Most of this development was focused in Johannesburg and in old East Rand mining communities such as Benoni, Boksburg, and Germiston, where the number of Africans employed in secondary industry soon exceeded the number working in the mines. Ominously for segregationists, a growing proportion of these workers were not migrants but permanently urbanized workers, living with wives and children.

The conflict between the imperatives of segregation and industrialization came to a head during World War II. With white workers off fighting and booming factories desperate for labour, the government of Prime Minister Jan Smuts suspended the operation of influx control on the Witwatersrand, unleashing a cataract of urbanization. Johannesburg's black population virtually doubled, to more than 400,000. With housing construction at a standstill, new arrivals were crammed into already overcrowded townships or into informal squatter camps thrown up wherever there was open space. Squalid, overcrowded conditions bred disease and vice, but they also spawned new forms of political consciousness and action, evident in bus boycotts in Alexandra, defiant squatter movements, and the rise of the militant African National Congress Youth League. A new wave of trade unionism spread among workers, culminating in the 1946 black mineworkers' strike.

APARTHEID

Enforcing the system. These transformations were not lost on white political leaders. On the contrary, the future of Johannesburg and other South African cities became the central issue in the 1948 national election. Jan Smuts' United Party, while defending its commitment to white supremacy, argued that complete segregation was chimerical and that some permanent African urbanization was an inevitable consequence of economic development. The National Party of Daniel F. Malan, in contrast, warned that whites were being "swamped" and called for a forceful restoration of the old order. Malan dubbed his policy "apartheid." Buoyed by a massive turnout of Afrikaner voters, the Nationalists secured a parliamentary majority, which they retained for the next 46 years. In that time they enacted a panoply of laws that specified where one could live, work, or attend school, all on the basis of race. The squatter camps that had sprung up around Johannesburg in the mid-1940s were bulldozed, as were established settlements that fell within areas now designated as "White." Sophiatown, home to many of South Africa's most celebrated writers and musicians, was removed starting in 1955. In its place rose a new white working-class suburb.

The National Party

The 1960s in Johannesburg were, depending on one's point of view, the best of times or the worst of times. With African oppositional movements banned, the National Party had a free hand to elaborate its vision of "grand apartheid." Between 1960 and 1980 close to four million blacks were forcibly relocated, including several hundred thousand from Johannesburg, to remote, desiccated ethnic "homelands." While careful not to impede the flow of labour to industry, state officials enforced pass laws with unprecedented rigour: in the quarter-century preceding the final repeal of influx control in 1986 the state prosecuted nearly 10 million pass offenses, an average of more than 1,000 per day. For whites, on the other hand, these years brought unimagined prosperity. Over the course of the 1960s the South African economy grew nearly 6 percent per annum, a rate exceeded only by Japan. Needless to say, most of the benefits accrued to whites, who rewarded the National Party with ever greater parliamentary majorities. (J.T.Ca.)

Apartheid's demise. Beneath apartheid's placid surface, however, lay seething discontent that exploded in Johannesburg in June 1976 when South African police fired on a group of Soweto student protestors. The shooting ignited a nationwide popular uprising that raged for months. Townships around Johannesburg exploded again in 1984, in response to the National Party's introduction of a new constitution that conceded limited franchise rights to Indians and Coloureds while excluding the African majority. Unrest continued to rage throughout the 1980s and periodically brought the economy of the Witwatersrand to a halt. This growing militancy helped propel South Africa's white rulers to the negotiating table, paving the way for the country's first democratic elections in 1994. Under the new ANC government, the Municipal Structures Act was passed into law in 1998, further integrating and regulating local government in Johannesburg and throughout the country. More controversial were plans to privatize large portions of Johannesburg's city services under a program called iGoli 2002, the goal of which was to make city services more responsive and accessible to all residents. Crime, traditionally a problem in Johannesburg, became particularly acute in the 1990s as violent crime drove businesses, including the Johannesburg Stock Exchange, from the city's centre into its safer northern suburbs.

BIBLIOGRAPHY. DIGBY RICCI (compiler and ed.), *Reef of Time: Johannesburg in Writing* (1986), contains a wide-ranging collection of recent and historical writings about the city. An illustrated discussion of the history, institutions, and culture of the city (though from a decidedly white perspective) can be found in *Johannesburg: One Hundred Years* (1986). A.P. CARTWRIGHT, *The Corner House: The Early History of Johannesburg* (1965), discusses history from the perspective of the great mining houses. CHARLES VAN ONSELEN, *Studies in the Social and Economic History of the Witwatersrand, 1886–1914*, 2 vol. (1982), a classic work, charts the social struggles unleashed by the opening of the Witwatersrand goldfields. ELLEN HELLMANN, *Rooiyard: A Sociological Survey of an Urban Native Slum Yard* (1948, reissued 1969), vividly describes the squalid, overcrowded conditions that prevailed among blacks in Johannesburg during and after World War II. For travelers, JON MURRAY, JEFF WILLIAMS, and RICHARD EVERIST, *South Africa, Lesotho and Swaziland*, 3rd ed. (1998), from the *Lonely Planet* series, gives practical information on Johannesburg. (J.T.Ca./Ed.)

Samuel Johnson

Samuel Johnson, English critic, biographer, essayist, poet, and lexicographer, once characterized literary biographies as "mournful narratives," and he believed that he lived "a life radically wretched." Yet his career can be seen as a literary success story of the sickly boy from the Midlands who by talent, tenacity, and intelligence became the foremost literary figure and the most formidable conversationalist of his time. For future generations, Johnson was synonymous with the later 18th century in England. The disparity between his circumstances and achievement gives his life its especial interest.

EARLY LIFE

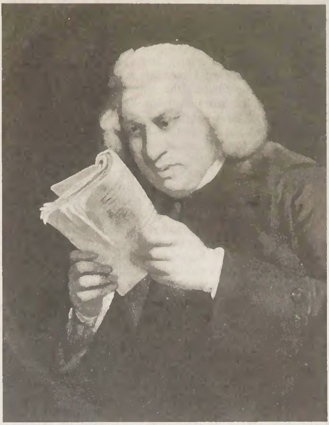
Physical afflictions

Samuel Johnson was born on Sept. 18, 1709, at Lichfield, Staffordshire, to Michael Johnson, a bookseller, and his wife, Sarah. From childhood he suffered from a number of physical afflictions and could have spoken, as had the 18th-century English poet Alexander Pope, of "this long disease my life." By his own account, he was born "almost dead," and he early contracted scrofula (tuberculosis of the lymphatic glands). Because of a popular belief that the sovereign's touch was able to cure scrofula, he was taken to London at the age of 30 months and touched by the queen, whose gold "touch piece" he kept about him for the rest of his life. This ineffective treatment was succeeded by various medical treatments that left him with disfiguring scars on his face and neck. He was nearly blind in his left eye and suffered from highly noticeable tics. Until he began to speak, new acquaintances sometimes took him for an idiot, as did the artist William Hogarth, who came to admire him greatly. Despite his many physical afflictions, Johnson was strong, vigorous, and, after a fashion, athletic. He liked to ride, walk, and swim, even in later life. He was tall and became huge. A few accounts bear witness to his physical strength—as well as his character—such as his hurling an insolent theatregoer together with his seat from the stage into the pit or his holding off would-be robbers until the arrival of the watch.

From his earliest years Johnson was recognized not only for his remarkable intelligence but also for his pride and indolence. In 1717 he entered grammar school in Lichfield. The master of the school, John Hunter, was a learned though brutal man who "never taught a boy in his life—he whipped and they learned." This regime instilled such terror in the young boy that even years later the resemblance of the poet Anna Seward to her grandfather Hunter caused him to tremble. At school he made two lifelong friends: Edmund Hector, later a surgeon, and John Taylor, future prebendary of Westminster and justice of the peace for Ashbourne. In 1726 Johnson visited his cousin, the urbane Reverend Cornelius Ford in Stourbridge, Worcestershire, who may have provided a model for him, though it was Ford's conviviality and scholarship rather than his dissipation (he is thought to be one of those depicted carousing in Hogarth's "A Midnight Modern Conversation" [1733]) that attracted Johnson.

Oxford

In 1728 Johnson entered Pembroke College, Oxford. He stayed only 13 months, until December 1729, because he lacked the funds to continue. Yet it proved an important year. Initially, he surprised his tutor William Jorden with his knowledge of obscure Latin poetry. More significantly, Johnson, who claimed to have been irreligious in adolescence, read a new book, William Law's *A Serious Call to a Devout and Holy Life,* which led him to make concern for his soul the polestar of his life. Despite the poverty and pride that caused him to leave, he retained great affection for Oxford. He would later say with reference to the poets of his college, "We were a nest of singing birds." In 1731, the year of his father's death, his first publication, a translation of Pope's "Messiah" into Latin, appeared in *A Miscellany of Poems,* along with the poetry of other



Samuel Johnson by Sir Joshua Reynolds, 1775. In a private collection.
By courtesy of Loren and Frances Rothschild

Oxford students. Pope was the leading poet of the age, and throughout most of his lifetime Johnson would comment on Pope's achievement in various writings.

Johnson left his books at Oxford, perhaps with thoughts of returning. In the following year the best he could do, despite his remarkable learning, was to become undermaster at Market Bosworth grammar school, a position made untenable by the overbearing and boorish Sir Wolstan Dixie, who controlled appointments. With only £20 inheritance from his father, Johnson left his position with the feeling that he was escaping prison. After failing in his quest for another teaching position, he joined his friend Hector in Birmingham. In 1732 or 1733 he published some essays in *The Birmingham Journal,* a newspaper, none of which have survived. Dictating to Hector, he translated into English Joachim Le Grand's translation of the Portuguese Jesuit Jerome Lobo's *A Voyage to Abyssinia,* an account of a Jesuit missionary expedition. Published in 1735, this work shows signs of the mature Johnson, such as his praise of Lobo, in the preface, for not attempting to present marvels: "He meets with no basilisks that destroy with their eyes, his crocodiles devour their prey without tears, and his cataracts fall from the rock without deafening the neighbouring inhabitants."

Marriage

In 1735 Johnson married Elizabeth Porter, a widow 20 years his senior. Convinced that his parents' marital unhappiness was caused by his mother's want of learning, he would not follow their example, choosing instead a woman whom he found both attractive and intelligent. His wife's marriage settlement enabled him to open a school in Edial, near Lichfield, the following year. One of his students, David Garrick, would become the greatest English actor of the age and a lifelong friend, though their friendship was not without its strains. It was with Garrick that some of the unflattering accounts of Johnson's wife originated, and his mimicry of the couple later became a favourite comic setpiece of his. While at Edial, Johnson began his historical tragedy *Irene,* which dramatizes the love of Sultan Mahomet (Mehmed II) for the lovely Irene, a Christian slave captured in Constantinople. The school soon proved a failure, and he and Garrick left for London in 1737. The schoolmaster in him survived, however, and can be detected at times in later works such as *A Dictionary of the English Language* and *The Lives of the Poets.*

THE GENTLEMAN'S MAGAZINE AND EARLY PUBLICATIONS

In 1738 Johnson began his long association with *The Gentleman's Magazine,* often considered the first modern magazine, which took a position of political neutrality. He soon contributed poetry and then prose, including panegyrics on Edward Cave, the magazine's proprietor, and another contributor, the learned Elizabeth Carter. Johnson intended to translate the Venetian Paolo Sarpi's *The History of the Council of Trent* but was forestalled by the coincidence of another Johnson at work on the same project. However, his biography of Sarpi, designed as a preface to that work, appeared in *The Gentleman's Magazine,* as did a number of his early biographies of European scholars, physicians, and British admirals.

In 1738 and 1739 he published a series of satiric works that attacked the government of Sir Robert Walpole and even the Hanoverian monarchy: *London* (his first major poem), *Marmor Norfolciense,* and *A Compleat Vindication of the Licensers of the Stage. London* is an "imitation" of the Roman satirist Juvenal's third satire. (A loose translation, an imitation applies the manner and topics of an earlier poet to contemporary conditions.) Thales, the poem's main speaker, bears some resemblance to the poet Richard Savage, of whom Johnson knew and with whom he may have become friendly at this time. Before he leaves the corrupt metropolis for Wales, Thales rails against the pervasive deterioration of London (and English) life, evident in such ills as masquerades, atheism, the excise tax, and the ability of foreign nations to offend against "English honour" with impunity. The most famous line in the poem (and the only one in capitals) is: "SLOW RISES WORTH, BY POVERTY DEPRESSED," which may be taken as Johnson's motto at this time. When the poem appeared anonymously in 1738, Pope was led to predict that its author would be "*déterré*" (unearthed). Pope undoubtedly approved of Johnson's politics along with admiring his poetry and tried unsuccessfully to arrange patronage for him. *Marmor Norfolciense* satirizes Walpole and the house of Hanover. Written, supposedly, by one "Probus Britanicus," a "true Briton," Johnson's fiction has a pro-government scholar find an ancient stone in Norfolk (Walpole's base) bearing obscure inscriptions that turn out to be innuendos against Walpole and the Hanover kings. *A Compleat Vindication of the Licensers of the Stage* is an ironic defense of the government's Stage Licensing Act of 1737 requiring the lord chamberlain's approval of all new plays, which in 1739 led to the prohibition of Henry Brooke's play *Gustavus Vasa* attacking the English monarch and his prime minister by Swedish analogy. The latter two works show the literary influence of the Irish writer Jonathan Swift.

Johnson at this time clearly supported the governmental opposition, which was composed of disaffected Whigs, Tories, Jacobites (those who continued their allegiance to the Stuart line of James II), and Nonjurors (those who refused to take either the oath of allegiance to the Hanover kings or the oath of abjuration of James II and the Stuarts). Despite claims to the contrary, Johnson was neither a Jacobite nor a Nonjuror.

Johnson's Toryism

His Toryism, which he sometimes expressed for shock value, was based upon his conviction that the Tories could be counted upon to support the Church of England as a state institution. When Johnson attacked Whiggism or defended Toryism (an ideology for him more than a practical politics, especially since Tories remained a minority throughout most of his lifetime), he always took an outsider's position. Later in life he expressed a high regard for Walpole.

In 1739 Johnson published a translation and annotation of the Swiss philosopher Jean-Pierre de Crousaz's *Commentary* on Pope's philosophical poem *An Essay on Man.* Although he was able to show that many of Crousaz's critical observations rested on a faulty French translation, Johnson often agreed with his judgment that some of Pope's philosophical and social ideas are marred by complacency. About this time Johnson tried again to obtain a position as a schoolteacher. His translations and magazine writings barely supported him; a letter to Cave is signed "impransus," signifying that he had gone without dinner. Despite his claim that "no man but a blockhead ever wrote except for money," he never made a hard bargain with a bookseller and often received relatively little payment, even for large projects. He also contradicted his assertion frequently by contributing prefaces and dedications to the books of friends without payment.

From 1741 to 1744 Johnson's most substantial contribution to *The Gentleman's Magazine* was a series of speeches purporting to represent the actual debates in the House of Commons. This undertaking was not without risk because reporting the proceedings of Parliament, which had long been prohibited, was actually punished since the spring of 1738. The series was dubbed "Debates in the Senate of Magna Lilliputia," and this Swiftian expedient gives the speeches satiric overtones. Their status was complicated by the fact that Johnson, who had visited the House of Commons only once, wrote the debates on the basis of scant information about the speakers' positions. Hence they were political fictions, though paradoxically they appeared to be fact masquerading as fiction. An essay of March 1740 comes close to describing Johnson's actual practice. "Mr. Gulliver" is represented as abridging the sometimes long-winded speeches and "may be said, like a judicious painter, to have marked the outlines, to have designed the principal proportions, and thrown in some characteristical strokes of each masterly hand, in these pictures of Lilliputian eloquence." Johnson later had misgivings about his role in writing speeches that were taken as authentic and may have stopped writing them for this reason. While Johnson's claim that he "took care that the Whig dogs should not have the best of it" has become notorious, Johnson's Walpole defends himself skillfully, and many of the debates seem even-handed.

In the early 1740s Johnson continued his strenuous work for *The Gentleman's Magazine;* collaborated with William Oldys, antiquary and editor, on a catalog of the great Harleian Library; helped Dr. Robert James, his Lichfield schoolfellow, with *A Medicinal Dictionary;* and issued proposals for an edition of Shakespeare. His *Miscellaneous Observations on the Tragedy of Macbeth* (1745), intended as a preliminary sample of his work, was his first significant Shakespeare criticism. In 1746 he wrote *The Plan of a Dictionary of the English Language* and signed a contract for *A Dictionary of the English Language.* His major publication of this period was *An Account of the Life of Mr. Richard Savage, Son of the Earl Rivers* (1744).

The *Life of Savage*

If, as Johnson claimed, the best biographies were written by those who had eaten and drunk and "lived in social intercourse" with their subjects, this was the most likely of his many biographies to succeed. The *Life* was widely admired by, among others, the painter Joshua Reynolds, and it was reviewed in translation by the French philosopher Denis Diderot. Although Johnson had few illusions about his self-publicizing friend's conduct and character, he nonetheless became his defender to a significant extent. Johnson's title supports Savage's claim to be the bastard son of a nobleman—a claim of which others have been highly skeptical—but his biography, in its mixture of pathos and satire, at once commemorates and criticizes Savage. Johnson thought that Savage's poverty cost society a great deal:

> On a bulk, in a cellar, or in a glasshouse among thieves and beggars, was to be found the author of *The Wanderer,* . . . the man whose remarks on life might have assisted the statesman, whose ideas of virtue might have enlightened the moralist, whose eloquence might have influenced senates, and whose delicacy might have polished courts.

Yet the conclusion leaves no doubt about Johnson's ultimate judgment: "negligence and irregularity, long continued, will make knowledge useless, wit ridiculous, and genius contemptible." If Johnson served as defense attorney throughout much of the biography, no prosecutor could have summed up the case against Savage more devastatingly.

MATURITY AND RECOGNITION

The Vanity of Human Wishes. In 1749 Johnson published *The Vanity of Human Wishes,* his most impressive poem as well as the first work published with his name. It is a panoramic survey of the futility of human pursuit of greatness and happiness. Like *London,* the poem is an imitation of one of Juvenal's satires, but it emphasizes the moral over the social and political themes of Juvenal. Some of the definitions Johnson later entered under "vanity" in his *Dictionary* suggest the range of meaning of his title, including "emptiness," "uncertainty," "fruitless desire, fruitless endeavour," "empty pleasure; vain pursuit; idle show; unsubstantial enjoyment; petty object of pride," and "arrogance." He portrays historical figures, mainly from England and continental Europe (Thomas Cardinal Wolsey, Charles XII of Sweden, the Persian king Xerxes I), alternating them with human types (the traveler, the rich man, the beauty, the scholar), often in juxtaposition with their opposites, to show that all are subject to the same disappointment of their desires. *The Vanity of Human Wishes* is imbued with the Old Testament message of Ecclesiastes that "all is vanity" and replaces Juvenal's Stoic virtues with the Christian virtue of "patience." The religion of the poem is universalized, deliberately referring to "heav'n" rather than a more specific sectarian conception of the deity, though the New Testament virtues of faith and charity ("love") play an important role in the conclusion, with "patience" substituting for hope. The poem surpasses any of Johnson's other poems in its richness of imagery and powerful conciseness.

The theatre. Johnson's connections to the theatre in these years included writing several prologues, one for Garrick's farce *Lethe* in 1740 and one for the opening of the Drury Lane theatre. Garrick, now its manager, returned the favours. Early in 1749 Johnson's play *Irene* was at last performed. Thanks to Garrick's production, which included expensive costumes, an excellent cast (including Garrick himself), and highly popular afterpieces for the last three performances, the tragedy ran a respectable nine nights. The audience objected to seeing the apostate Greek Christian Irene strangled by Sultan Mahomet—an innovation of Garrick's—and the murder was performed offstage thereafter. *Irene* is Johnson's least appealing major work, and he is reported to have said when hearing someone read it aloud, "I thought it had been better."

Performance of *Irene*

From *The Rambler* to *The Adventurer*. With *The Rambler* (1750–52), a twice-weekly periodical, Johnson entered upon the most successful decade of his career. He wrote over 200 numbers, and stories abound of his finishing an essay while the printer's boy waited at the door; in his last essay he confessed to "the anxious employment of a periodical writer." The essays cover a wide range of subjects. A large number of them appropriately stress daily realities; others are devoted to literature, including criticism and the theme of authorship (particularly the early ones, driven by the writer's consciousness of his own undertaking) and to literary forms, such as the novel and biography, that had not received much examination. Whatever their topic, Johnson intended his essays to "inculcate wisdom or piety" in conformity with Christianity. Thus, a typical essay (No. 32) on "the art of bearing calamities" counsels patience, as had *The Vanity of Human Wishes,* and insists that "The cure for the greatest part of human misery is not radical, but palliative." In tone these essays are far more serious than those of his most important predecessor, Joseph Addison, published in *The Spectator* (1711–12; 1714). The 20th-century critic Walter Jackson Bate finds them "saturated with thought to an extent unexceeded by any other writer of English prose since Francis Bacon." Johnson himself ranked them highly among his achievements, commenting "My other works are wine and water; but my *Rambler* is pure wine." Although *The Rambler* may have sold only 500 copies an issue on its first appearance—in his last number he claimed he had "never been much a favourite of the public"—it was widely reprinted in provincial newspapers and sold well in later editions.

Johnson's *Rambler* series also was admired by his wife Elizabeth, who praised its author by saying, "I thought very well of you before this; but I did not imagine you could have written any thing equal to this." She died on March 17, 1752, just three days after the publication of its last number. In her later years "Tetty" frequently lived away from him in Hampstead. Signs of marital tensions may be glimpsed in surviving letters and in Johnson's prayers, which were published after his death. He wrote a sermon for her funeral that praises her submissive piety—her "exact and regular" devotions—as well as her charitable disposition.

A diary entry suggests that a year after Elizabeth's death Johnson was seeking a new wife "without any derogation from dear Tetty's memory." The one he most probably had in mind was the pious Hill Boothby, to whom he wrote with some frequency in the years immediately following this resolve. Three dozen of her letters to him, rarely quoted by biographers, are in print. The relationship, however, came to an end with her death in 1756.

Hill Boothby

During the course of one year starting in March 1753, Johnson contributed 29 essays to his friend John Hawkesworth's periodical *The Adventurer,* written in imitation of *The Rambler.* Johnson purposely (and ineffectively) lightened his style in order to hide his authorship. He wanted his essays unrecognized, for he had given them to Dr. Richard Bathurst, the friend whom he said he loved more than any other, to sell as his own, but he confessed his part to the persistent Hill Boothby.

The *Dictionary*. *A Dictionary of the English Language* was published in two volumes in 1755, six years later than planned but remarkably quickly for so extensive an undertaking. The degree of Master of Arts, conferred on him by the University of Oxford for his *Rambler* essays and the *Dictionary,* was proudly noted on the title page. Johnson henceforth would be known in familiar 18th-century style as "Dictionary Johnson" or "The Rambler." There had been earlier English dictionaries, but none on the scale of Johnson's. In addition to giving etymologies, not the strong point of Johnson and his contemporaries, and definitions, in which he excelled, Johnson illustrated usage with quotations drawn almost entirely from writing from the Elizabethan period to his own time, though few living authors were quoted (the novelists Samuel Richardson and Charlotte Lennox, Garrick, Reynolds, and Johnson himself among them). His preface boldly asserts that the "chief glory of every people arises from its authors," and his book (the phrase he always used for it) was his own claim to be ranked among them. He was pleased that what took the French Academy 40 years to perform for their language was accomplished by one Englishman in 9 years. It may have been his desire to fix the language by his work, yet he realized that languages do not follow prescription but are continually changing. Johnson did not work systematically from a word list but marked up the books he read for copying. Thus it is no surprise that some earlier dictionaries contain more words and that Johnson's has striking omissions ("literary" for one). Yet his definitions were a great improvement over those of his predecessors, and his illustrations from writers since the Elizabethan Age form an anthology and established a canon. Because he insisted not only on correct usage but also on morality and piety, the illustrations of words often come from sermons and conduct books as well as from a range of literature. The skeptical philosopher Thomas Hobbes and the writer Bernard de Mandeville, who praised the public benefits of brothels, were excluded on moral grounds, and in the *Plan* for the *Dictionary* Johnson explains that the inclusion of a writer could be taken as an invitation to read his work.

Johnson had been persuaded to address his *Plan* to the Earl of Chesterfield as his patron, but his appeal had been met with years of neglect. Johnson's defensive pride was awakened when the nobleman, learning of the impending publication of the *Dictionary,* praised it in two essays in *The World,* a weekly paper of entertainment. His letter to Chesterfield is often taken as sounding "the death-knell of patronage," which it did not. But it did assert the dignity of the author.

Lord Chesterfield

> Is not a patron, my Lord, one who looks with unconcern on a man struggling for life in the water, and, when he has reached ground, encumbers him with help. The notice which you have been pleased to take of my labours, had it been early, had been kind; but it has been delayed till I am indifferent and cannot enjoy it, till I am solitary and cannot impart it, till I am known, and do not want it.

The *Dictionary* defines "patron" as "one who countenances, supports, or protects. Commonly a wretch who supports with insolence, and is paid with flattery."

The personal quality of the *Dictionary*

Johnson's *Dictionary* is both a personal work and deeply responsive to dictionary tradition. It has often been noted that its definitions suggest autobiography or prejudice. "Lich," for example, is ornamented with a Latin salute to his native Lichfield; "Whig" is defined as "the name of a faction"; "excise" is "a hateful tax levied upon commodities, and adjudged not by the common judges of property, but wretches hired by those to whom excise is paid." Yet the *Dictionary* is a personal work above all in its choice of authors and of illustrative selections, which give the whole the aspect of both an encyclopaedia and a conduct book. Even though Johnson defined "lexicographer" as "a writer of dictionaries; a harmless drudge," the drudgery of the *Dictionary* fell into the decade of Johnson's most important writing and must be seen in part as enabling it. The payment for the *Dictionary* amounted to relatively little after deductions were made for his six amanuenses and his own expenses. He left his house in Gough Square (now the most famous of Johnson museums) for smaller lodgings in 1759, ending the major decade of his literary activity famous and poor.

The Literary Magazine. From 1756 onward Johnson wrote harsh criticism and satire of England's policy in the Seven Years' War (1756–63) fought against France (and others) in North America, Europe, and India. This work appeared initially in a new journal he was editing, *The Literary Magazine,* where he also published his biography of the Prussian king, Frederick II (the Great). He also contributed important book reviews when reviewing was still in its infancy. His bitingly sardonic dissection of a dilettantish and complacent study of the nature of evil and of human suffering, *A Free Enquiry into the Nature and Origin of Evil,* by the theological writer Soame Jenyns, may well be the best review in English during the 18th century:

> This author and Pope perhaps never saw the miseries which they imagine thus easy to be borne. The poor indeed are insensible of many little vexations which sometimes embitter the possessions and pollute the enjoyments of the rich. They are not pained by casual incivility, or mortified by the mutilation of a compliment; but this happiness is like that of a malefactor who ceases to feel the cords that bind him when the pincers are tearing his flesh.

Johnson intended to bring out still another journal, *The Annals of Literature,* that would consider foreign as well as domestic literature. Although this project came to nothing, it helps to demonstrate that a tendency to regard Johnson as the literary incarnation of John Bull, Englishman, ignores his genuine interest in other cultures.

The Idler. Johnson's busiest decade was concluded with yet another series of essays, called *The Idler.* Lighter in tone and style than those of *The Rambler,* its 104 essays appeared from 1758 to 1760 in a weekly newspaper, *The Universal Chronicle.* While not admired as greatly as *The Rambler,* Johnson's last essay series contained many impressive numbers, such as No. 84, in which he praised autobiography over biography and drew his self-portrait as "Mr. Sober," a consummate idler. The original No. 22, his account of an old vulture explaining to her offspring man's propensities as a killer and concluding that man more than any other animal is "a friend to vultures," was considered too strong to be included in the collected editions.

Rasselas. Johnson's essays included numerous short fictions, but his only long fiction is *Rasselas* (originally published as *The Prince of Abissinia: A Tale*), which he wrote in 1759, during the evenings of a single week, in order to be able to pay for the funeral of his mother. This "Oriental tale," a popular form at the time, explores and exposes the futility of the pursuit of happiness, a theme that links it to *The Vanity of Human Wishes.* Prince Rasselas, weary of life in the Happy Valley, where ironically all are dissatisfied, escapes with his sister and the widely traveled poet Imlac to experience the world and make a thoughtful "choice of life." Yet their journey is filled with disappointment and disillusionment. They examine the lives of men in a wide range of occupations and modes of life in both urban and rural settings—rulers and shepherds, philosophers, scholars, an astronomer, and a hermit. They discover that all occupations fail to bring satisfaction. Rulers are deposed. The shepherds exist in grubby ignorance, not pastoral ease. The Stoic's philosophy proves hollow when he experiences personal loss. The hermit, miserable in his solitude, leaves his cell for Cairo. In his "conclusion in which nothing is concluded," Johnson satirizes the wish-fulfilling daydreams in which all indulge. His major characters resolve to substitute the "choice of eternity" for the "choice of life," and to return to Abyssinia (but not the Happy Valley) on their circular journey.

Pension

Johnson never again had to write in order to raise funds. In 1762 he was awarded a pension of £300 a year, "not," as Lord Bute, the prime minister, told him, "given you for anything you are to do, but for what you have done." This in all likelihood meant not only his literary accomplishments but also his opposition to the Seven Years' War, which the new king, George III, and his prime minister had also opposed. Although in his *Dictionary* Johnson had added to his definition of "pension," "In England it is generally understood to mean pay given to a state hireling for treason to his country," he believed that he could accept his with a clear conscience.

Friendships and household. In 1763 Johnson met the 22-year-old James Boswell, who would go on to make him the subject of the best-known and most highly regarded biography in English. The first meeting with this libertine son of a Scottish laird and judge was not auspicious, but Johnson quickly came to appreciate the ingratiating and impulsive young man. Boswell kept detailed journals, published only in the 20th century, which provided the basis for his biography of Johnson and also form his own autobiography.

The Club

Johnson participated actively in clubs. In 1764 he and his close friend Sir Joshua Reynolds founded The Club (later known as The Literary Club), which became famous for the distinction of its members. The original nine members included the politician Edmund Burke, the playwright Oliver Goldsmith, and Sir John Hawkins, the historian of music whom Johnson was to call "unclubable." Boswell, whose 1768 account of the Corsican struggle against Genoese rule and its revolutionary leader, General Pasquale Paoli, earned him a reputation throughout Europe, was admitted in 1773. Other members elected later included Garrick, the historian Edward Gibbon, the dramatist Richard Brinsley Sheridan, the economist and moral philosopher Adam Smith, and the Orientalist Sir William Jones. In 1749 Johnson had been one of 10 members of the Ivy Lane Club, and the year before his death he founded The Essex Head Club. These clubs, at which he often "talked for victory," provided the conversation and society he desired and kept him from the loneliness and insomnia that he faced at home.

This is not to say that his house was empty after the death of his wife. He had living with him at various times Anna Williams, a blind poet; Elizabeth Desmoulins, the daughter of his godfather Dr. Samuel Swynfen, and her daughter; Poll Carmichael, probably a former prostitute; "Dr." Robert Levett, a medical practitioner among the poor; Francis Barber, Johnson's black servant, whom he treated in many ways like a son and made his heir; and Barber's wife Betsy. They were at once recipients of Johnson's charity and providers of company, but the relationship among them was not always amicable. In a letter of 1778 Johnson says, "We have tolerable concord at home, but no love. Williams hates everybody; Levett hates Desmoulins, and does not love Williams; Desmoulins hates them both; Poll loves none of them."

Henry Thrale

In 1765 Johnson established a friendship that soon enabled him to call another place "home." Henry Thrale, a wealthy brewer and member of Parliament for Southwark, and his lively and intelligent wife, Hester, opened their country house at Streatham to him and invited him on trips to Wales and, in 1775, to France, his only tour outside Great Britain. Their friendship and hospitality gave the 56-year-old Johnson a new interest in life. Following her husband's death in 1781 and her marriage to her children's music master, Gabriel Piozzi, Hester Thrale's and Johnson's close friendship came to an end. His letters to Mrs. Thrale, remarkable for their range and intimacy, helped make him one of the great English letter writers.

The edition of Shakespeare. The pension Johnson had received in 1762 had freed him from the necessity of writing for a living, but it had not released him from his obligation to complete the Shakespeare edition, for which he had taken money from subscribers. His long delay in bringing that project to fruition provoked some satiric notice from the poet Charles Churchill:

> He for subscribers baits his hook,
> And takes their cash—but where's the book?

The edition finally appeared in eight volumes in 1765. Johnson edited and annotated the text and wrote a preface, which is his greatest work of literary criticism. As editor and annotator he sought to establish the text, freed from later corruptions, and to explain diction that by then had become obsolete and obscure. He remarks dryly of an annotation from an edition of Shakespeare's works by his contemporary William Warburton, the bishop of Gloucester, saying that "it explains what no reader has found difficult, and, I think, explains it wrong." Johnson, however, was fully aware that perfection was impossible to achieve, since "every work of this kind is by its nature deficient," and responded to Pope's complaints of "the dull duty of an editor" that "conjectural criticism demands more than humanity possesses." Johnson's approach was to immerse himself in the books Shakespeare had read—his extensive reading for his *Dictionary* eased this task—and to examine the early editions as well as those of his 18th-century predecessors. His annotations are often shrewd, though his admiration reveals at times different concerns from those of some of his contemporaries and of later scholars. While he joins in the general praise of Sir John Falstaff in *Henry IV, Part I,* he gives special recognition to a scene from *Henry VIII* in which the sick Queen Katharine hears of the death of Cardinal Wolsey:

> This scene is above any other part of Shakespeare's tragedies, and perhaps above any scene of any other poet, tender and pathetick, without gods, or furies, or poisons, or precipices, without the help of romantick circumstances, without improbable sallies of poetical lamentation, and without any throes of tumultuous misery.

The "Preface"

In his "Preface" Johnson addressed several critical issues. For one, he vigorously defends Shakespeare against charges of failing to adhere to the Neoclassical doctrine of the dramatic unities of time, place, and action. Johnson alertly observes that time and place are subservient to the mind: since the audience does not confound stage action with reality, it has no trouble with a shift in scene from Rome to Alexandria. Some critics had made similar points before, but Johnson's defense was decisive. He also questions the need for purity of dramatic genre. In defending Shakespearian tragicomedy against detractors, he asserts that "there is always an appeal open from criticism to nature." Echoing Hamlet, Johnson claims that Shakespeare merits praise, above all, as "the poet of nature; the poet that holds up to his readers a faithful mirror of manners and of life." He goes on to say that "in the writings of other poets a character is too often an individual: in those of Shakespeare it is commonly a species" and that "Shakespeare has no heroes; his scenes are occupied only by men." These comments inveigh against the rigid notions of decorum upheld by critics, such as Voltaire, who would not allow kings to be drunkards or senators to be buffoons. Johnson's concern for "general nature" means that he is not much interested in accidental traits of a character, such as the "Romanness" of Julius Caesar or Brutus, but in traits that are common to all humanity.

Dr. Johnson. In 1765 Johnson received an honorary Doctor of Laws degree from Trinity College, Dublin, and 10 years later he was awarded the Doctor of Civil Laws from the University of Oxford. He never referred to himself as Dr. Johnson, though a number of his contemporaries did, and Boswell's consistent use of the title in *The Life of Samuel Johnson, LL.D.* made it popular. The completion of the Shakespeare edition left Johnson free to write by choice, and one such choice was his secret collaboration with Robert Chambers, professor of English law at the University of Oxford from 1766 to 1773. While it is difficult to determine just how much of Chambers' lectures Johnson may have written, his help was clearly substantial, and the skilled editor was valued by the dilatory professor.

Political pamphlets. In the early 1770s Johnson wrote a series of political pamphlets supporting positions favourable to the government but in keeping with his own views. These have often appeared reactionary to posterity but are worth considering on their own terms. *The False Alarm* (1770) supported the resolution of the House of Commons not to readmit one of its members, the scandalous John Wilkes, who had been found guilty of libel. The pamphlet ridiculed those who thought the case precipitated a constitutional crisis. *Thoughts on the Late Transactions Respecting Falkland's Islands* (1771) argued against a war with Spain over who should become "the undisputed lords of tempest-beaten barrenness." This pamphlet, his most admired and least attacked, disputes the "feudal gabble" of the Earl of Chatham and the complaints of the pseudonymous political controversialist who wrote the "Junius" letters. *The Patriot* (1774) was designed to influence an upcoming election. Johnson had become disillusioned in the 1740s with those members of the political opposition who attacked the government on "patriotic" grounds only to behave similarly once in power. This essay examines expressions of false patriotism and includes in that category justifications of "the ridiculous claims of American usurpation," the subject of his longest tract, *Taxation No Tyranny* (1775). The title summarizes his position opposing the American Continental Congress, which in 1774 had passed resolutions against taxation by England, perceived as oppression, especially since the colonies had no representation in parliament. Johnson argues that the colonists had not been denied representation but rather had willingly left the country where they had votes, that England had expended vast sums on the colonies, and that they were rightly required to support the home country. The tract became notorious in the colonies, contributing considerably to the caricature of Johnson the arch-Tory. Yet this view is too simplistic. His rhetorical question to the colonists, "How is it that we hear the loudest yelps for liberty from the drivers of Negroes?" can be traced in large part to a principled and consistent stance against colonial oppression.

Taxation No Tyranny

Journey to the Hebrides. In 1773 Johnson set forth on a journey to the Hebrides. Given his age, ailments, and purported opinion of the Scots, Johnson may have seemed a highly unlikely traveler to this distant region, but in the opening pages of his *A Journey to the Western Islands of Scotland* (1775) he confessed to a longstanding desire to make the trip and the inducement of having Boswell as his companion. He was propelled by a curiosity to see strange places and study modes of life unfamiliar to him. His book, a superb contribution to 18th-century travel literature, combines historical information with what would now be considered sociological and anthropological observations about the lives of common people. (Boswell's complementary narrative of their journey, *The Journal of a Tour to the Hebrides,* with its rich store of Johnson's conversation, was published only in 1785, the year after Johnson's death.)

The Lives of the Poets. Johnson's last great work, *Prefaces, Biographical and Critical, to the Works of the English Poets* (conventionally known as *The Lives of the Poets*), was conceived modestly as short prefatory notices to an edition of English poetry. When Johnson was approached by some London booksellers in 1777 to write what he thought of as "little Lives, and little Prefaces, to a little edition of the English Poets," he readily agreed. He loved anecdote and "the biographical part" of literature best of all. The project, however, expanded in scope; Johnson's prefaces alone filled the first 10 volumes (1779–81), and the poetry grew to 56 volumes. Johnson was angered by the appearance of his name on the spines because he had neither "recommended" nor "revised" these poets, except for adding Isaac Watts, Sir Richard Blackmore, John Pomfret, Thomas Yalden, and James Thomson to the list.

While modern scholarly biographies give more accurate accounts of their subjects' lives, Johnson's work is of enduring value. He often sympathized with his subjects, yet he did not flinch from truth: "if a man is to write *A Panegyric,* he may keep vices out of sight; but if he professes to write *A Life,* he must represent it really as it was." Johnson steadily reminds his readers that even the greatest of poets are human. The lives are ordered chronologically by date of death, not birth, and range in length from a few pages to an entire volume. Among the major lives are those of Abraham Cowley, John Milton, John Dryden, Joseph Addison, and Alexander Pope; some of the minor ones, such as those of William Collins and William Shenstone, are little gems. Johnson's personal dislike of some of the poets whose lives he wrote, such as John Milton and Thomas Gray, has been used as a basis for arguing that he was prejudiced against their poetry, but too much has been made of this. His opinions of a poet and his work diverge at times as, for example, in the case of Collins. Johnson liked the man but disapproved of his poetic manner: "he puts his words out of the common order, seeming to think, with some later candidates for fame, that not to write prose is certainly to write poetry." He was justly proud of *The Life of Cowley,* especially of its lengthy discussion of the 17th-century Metaphysical poets, of whom Cowley may be considered the last representative. *The Life of Pope* is at once the longest and best. Pope's life and career were fresh enough and public enough to provide ample biographical material. Johnson found Pope's poetry highly congenial. His moving, unsentimental account of Pope's life is sensitive to his physical sufferings and yet unwilling to accept them as an excuse. His riposte to Pope's detractors, such as the poet Joseph Warton, is vigorous and memorable: "It is surely superfluous to answer the question that has once been asked, whether Pope was a poet? otherwise than by asking, in return, if Pope be not a poet, where is poetry to be found?" Yet in his masterly comparison of Pope and Dryden he acknowledges Dryden as the greater poet.

The lives of Cowley and Pope

Johnson divided his biographies into three distinct parts: a narrative of the poet's life, a presentation of his character (summarized traits), and a critical assessment of his main poems. He adopted this method not because he failed to perceive relationships between a poet's life and his works but because he did not think that a good poet was necessarily a good man. His method allowed him to make use of his recognition that "a manifest and striking contrariety between the life of an author and his writings" can exist and to assign different purposes to his analysis of his subjects' lives and their poetry. Johnson expressed a hope that the biographical parts of his lives were composed "in such a manner, as may tend to the promotion of Piety," and his moral intent is borne out in his readiness to chastise failings and to commend virtue. Johnson responded most favourably to the works of poets from Dryden to Pope and was skeptical of those produced in his own generation, including the poetry of Gray, Collins, and Shenstone, though he admired Gray's *An Elegy Written in a Country Church Yard.*

Later poetry. After his two Juvenalian satires Johnson wrote no ambitious poetry, but his later work includes a handful of memorable poems. His short elegy "On the Death of Dr. Robert Levet," who had practiced medicine among the London poor, portrays humans as "condemned to hope's delusive mine," existing like prisoners at hard labour, and he praises Levett, "obscurely wise, and coarsely kind," for having made good use of his "single talent." Other poems of note are a comic one for Mrs. Thrale's 35th birthday, which mingles affection and light satire in a way reminiscent of Swift's poems to Stella, and "A Short Song of Congratulation" to a callow heir, which is at once lively and wry in its ironic treatment of the carpe diem theme.

LAST YEARS

Throughout much of his adult life Johnson suffered from physical ailments as well as depression ("melancholy"). After the loss of two friends, Henry Thrale in 1781 and Robert Levett in 1782, and the conclusion of *The Lives of the Poets,* his health deteriorated. Above all, his chronic bronchitis and "dropsy" (edema), a swelling of his legs and feet, caused great discomfort. In 1783 he suffered a stroke. His last year was made still bleaker by his break with Mrs. Thrale over her remarriage. He compared himself at one point to those from whom confessions were extorted by the placement of heavy stones upon their chests. Yet he insisted on fighting: "I will be conquered; I will not capitulate." A profoundly devout Anglican, Johnson was in dread at the prospect of death and judgment, for he feared damnation. Yet in the winter of 1784, following a day of prayer after which his edema spontaneously disappeared, he entered into a previously unknown state of serenity. He accepted this release from illness as a sign that he might be saved after all and referred to it as a "late conversion." He died on Dec. 13, 1784, and was buried in Westminster Abbey.

A "late conversion"

ACHIEVEMENT AND REPUTATION

Johnson is well remembered for his aphorisms, which contributed to his becoming, after Shakespeare, the most frequently quoted of English writers. Many of these are recorded in Boswell's *The Life of Samuel Johnson, LL.D.,* including his famous assertion: "Patriotism is the last refuge of a scoundrel" and his admonition: "Clear your mind of cant." Others appear in his own writings, including: "Marriage has many pains, but celibacy has no pleasures." He possessed the gift of contracting "the great rules of life into short sentences."

Johnson's criticism is, perhaps, the most significant part of his writings. His assessment of Dryden's critical works holds good for his own: "the criticism of Dryden is the criticism of a poet; not a dull collection of theorems, nor a rude detection of faults, which perhaps the censorer was not able to have committed; but a gay and vigorous dissertation, where delight is mingled with instruction, and where the author proves his right of judgment by his power of performance." Although some have spoken of Johnson as a "literary dictator," he rejected the role for himself and in general spoke against the notion of enforcing precepts. A mere listing of the critical terms he used, which are not idiosyncratic but widely shared, would do little to explain his criticism. He required both truthfulness of representation and morality. Many have praised his common sense, but the flexibility and coherence of his response to literature is even more important.

Religion was central to Johnson's understanding of literature and of the moral life generally. His personal uneasiness about religion seems traceable to an orthodox fear that he might be among the damned. He saw himself as someone who did not practice what he preached and lived in dread that he would be, in the words of St. Paul, a castaway. His watch bore in Greek the biblical text, "The night cometh," a reminder of death and work left undone.

The role of religion

Johnson is more complex than he is often taken to be. His wide range of interests included science and manufacturing processes, and his knowledge seemed encyclopaedic. Although his late political tracts in defense of the government are antidemocratic, Johnson combined a high regard for monarchy with a low opinion of most kings. He frequently expressed minority or unpopular views, such as his principled stands against slavery, colonialism, and mistreatment of indigenous peoples. He also urged better treatment of prisoners of war, prostitutes, and the poor generally, and he once tried to save a convicted forger from the gallows.

If, as has often been claimed—largely because of Boswell's biography—we know Johnson as we know few other people in history (or few other characters in literature), we know him primarily as a man who overcame a host of difficulties to become the leading scholar and writer of his age. His imposing scope made him what might now be called a public intellectual. In the 19th century the interest in Johnson centred on his personality, the subject of Boswell's biography. In the 20th century his writings regained their rightful prominence.

MAJOR WORKS

VERSE: *London: A Poem* (1738); *The Vanity of Human Wishes* (1749); *Irene: A Tragedy* (1749).

PROSE: *Marmor Norfolciense* (1739); *A Compleat Vindication of the Licensers of the Stage* (1739); *An Account of the Life of Mr. Richard Savage, Son of the Earl Rivers* (1744); *Miscellaneous Observations on the Tragedy of Macbeth* (1745); *The Plan of a Dictionary of the English Language* (1747); *The Rambler*, 208 numbers (1750–52); *A Dictionary of the English Language*, 2 vol. (1755); *The Prince of Abissinia: A Tale*, 2 vol. (1759), better known as *Rasselas; The Idler* (1758–60); *The Plays of William Shakespeare*, 8 vol. (1765); *The False Alarm* (1770); *Thoughts on the Late Transactions Respecting Falkland's Islands* (1771); *The Patriot* (1774); *Taxation No Tyranny* (1775); *A Journey to the Western Islands of Scotland* (1775); *Prefaces, Biographical and Critical, to the Works of the English Poets*, 10 vol. (1779–81), best known as *The Lives of the Poets; Prayers and Meditations*, ed. by GEORGE STRAHAN (1785).

EDITIONS: *The Works of Samuel Johnson*, 11 vol. (1825, reprinted as *Dr. Johnson's Works*, 11 vol., 1970), was standard only because it was most cited. It is being replaced by *The Yale Edition of the Works of Samuel Johnson* (1958–). Until superseded, *The Lives of the English Poets*, ed. GEORGE BIRKBECK HILL, 3 vol. (1905, reissued 1967), is the standard edition; although J.P. HARDY (ed.), *Johnson's Lives of the Poets: A Selection* (1971), is preferable for the complete major lives he chooses. *Life of Savage*, ed. by CLARENCE TRACY (1971), is the best edition of this work. For the correspondence, *The Letters of Samuel Johnson*, ed. by BRUCE REDFORD, 5 vol. (1992–94), known as the *Hyde Edition*, is now the standard. The best one-volume anthology is DONALD GREENE (ed.), *Samuel Johnson* (1984). J.D. FLEEMAN (ed.), *The Complete English Poems* (1971, reprinted 1982), is a useful collection.

BIBLIOGRAPHY. The standard bibliography is WILLIAM P. COURTNEY and DAVID NICHOL SMITH, *A Bibliography of Samuel Johnson* (1915, reprinted 1984); it is supplemented by R.W. CHAPMAN and ALLEN T. HAZEN, "Johnsonian Bibliography: A Supplement to Courtney," in OXFORD BIBLIOGRAPHICAL SOCIETY, *Proceedings and Papers*, vol. 5, pt. 3, p. 117–166 (1938, reprinted 1984). JAMES L. CLIFFORD and DONALD J. GREENE, *Samuel Johnson: A Survey and Bibliography of Critical Studies* (1970), contains around 4,000 items; it is supplemented by DONALD GREENE and JOHN A. VANCE, *A Bibliography of Johnsonian Studies, 1970–1985* (1987). Recent work can be found in *The Age of Johnson* (annual); and *Johnsonian News Letter* (quarterly). Other bibliographic aids include J.D. FLEEMAN, *A Preliminary Handlist of Documents and Manuscripts of Samuel Johnson* (1967); DONALD GREENE, *Samuel Johnson's Library: An Annotated Guide* (1975); and HELEN LOUISE MCGUFFIE, *Samuel Johnson in the British Press, 1749–1784: A Chronological Checklist* (1976).

The number of biographical and critical studies of Johnson is enormous and rapidly growing. Among contemporary accounts are JAMES BOSWELL, *The Journal of a Tour to the Hebrides* (1785), available in many later editions, and *The Life of Samuel Johnson, LL.D.* (1791), also available in numerous editions—the best one is ed. by GEORGE BIRKBECK HILL, rev. and enlarged by L.F. POWELL, 6 vol. (1934–50, reissued 1979); *The Yale Editions of the Private Papers of James Boswell*, ed. by FREDERICK A. POTTLE, 15 vol. (1950–81), particularly vol. 2, *The Correspondence and Other Papers of James Boswell Relating to the Making of the Life of Johnson*, ed. by MARSHALL WAINGROW (1968); JAMES BOSWELL, *James Boswell's Life of Johnson*, ed. by MARSHALL WAINGROW (1994–), an edition of the original manuscript; JOHN HAWKINS, *The Life of Samuel Johnson, LL.D.* (1787, reprinted 1974); HESTER LYNCH PIOZZI, *Anecdotes of the Late Samuel Johnson, LL.D.* (1786, reprinted 1974); ARTHUR MURPHY, *An Essay on the Life and Genius of Samuel Johnson, LL.D.* (1792, reprinted 1970); and GEORGE BIRKBECK HILL (ed.), *Johnsonian Miscellanies*, 2 vol. (1897, reprinted 1966), with many anecdotes as well as selections from the early biographies. O M BRACK, JR., and ROBERT E. KELLEY (eds.), *The Early Biographies of Samuel Johnson* (1974), reprints fourteen lives. Also of interest are ROBERT E. KELLEY and O M BRACK, JR., *Samuel Johnson's Early Biographers* (1971); and MARY HYDE, *The Impossible Friendship: Boswell and Mrs. Thrale* (1972). Two famous essays by THOMAS BABINGTON MACAULAY, one a review of J.W. Croker's edition of Boswell's *Life*, "Samuel Johnson," *The Edinburgh Review* (September 1831), also available in his *Critical and Historical Essays, Contributed to The Edinburgh Review*, vol. 1 (1843), pp. 353–407, and the other, "Johnson, Samuel," in *The Encyclopædia Britannica*, 8th ed., vol. 12 (1856), pp. 793–803, set the tone for 19th-century denigration of Johnson's work. Johnson's genealogy and other finely sifted information has been assembled by ALEYN LYELL READE, *Johnsonian Gleanings*, 11 vol. (1909–52, reprinted 11 vol. in 10, 1968). Early critical reception is collected in JAMES T. BOULTON (ed.), *Johnson: The Critical Heritage* (1971).

Twentieth-century biographies and biographical studies include JAMES L. CLIFFORD, *Young Sam Johnson* (1955, reissued 1981; also published as *Young Samuel Johnson*, 1955), and *Dictionary Johnson: Samuel Johnson's Middle Years* (1979); W. JACKSON BATE, *Samuel Johnson* (1977); THOMAS KAMINSKI, *The Early Career of Samuel Johnson* (1987); JOHN WAIN, *Samuel Johnson*, rev. ed. (1988); ROBERT DEMARIA, JR., *The Life of Samuel Johnson* (1993); and RICHARD HOLMES, *Dr. Johnson & Mr. Savage* (1993).

Among general studies are BERTRAND H. BRONSON, *Johnson Agonistes, and Other Essays* (1946, reissued 1965); W. JACKSON BATE, *The Achievement of Samuel Johnson* (1955, reissued 1978); DONALD GREENE, *Samuel Johnson*, updated ed. (1989); PAUL FUSSELL, *Samuel Johnson and the Life of Writing* (1971, reissued 1986); J.P. HARDY, *Samuel Johnson: A Critical Study* (1979); ISOBEL GRUNDY, *Samuel Johnson and the Scale of Greatness* (1986); and PAT ROGERS, *Johnson* (1993), a brief work.

Special studies include MAURICE J. QUINLAN, *Samuel Johnson: A Layman's Religion* (1964); CHESTER F. CHAPIN, *The Religious Thought of Samuel Johnson* (1968); JAMES GRAY, *Johnson's Sermons: A Study* (1972); CHARLES E. PIERCE, JR., *The Religious Life of Samuel Johnson* (1983); NICHOLAS HUDSON, *Samuel Johnson and Eighteenth-Century Thought* (1988); PAUL K. ALKON, *Samuel Johnson and Moral Discipline* (1967); JOHN WILTSHIRE, *Samuel Johnson in the Medical World* (1991); GLORIA SYBIL GROSS, *This Invisible Riot of the Mind: Samuel Johnson's Psychological Theory* (1992); RICHARD B. SCHWARTZ, *Samuel Johnson and the New Science* (1971), and *Daily Life in Johnson's London* (1983); DONALD GREENE, *The Politics of Samuel Johnson*, 2nd ed. (1990); JOHN CANNON, *Samuel Johnson and the Politics of Hanoverian England* (1994); J.C.D. CLARK, *Samuel Johnson: Literature, Religion, and English Cultural Politics from the Restoration to Romanticism* (1994); ROBERT FOLKENFLIK, *Samuel Johnson, Biographer* (1978); DAVID WHEELER (ed.), *Domestick Privacies: Samuel Johnson and the Art of Biography* (1987); JOHN A. VANCE, *Samuel Johnson and the Sense of History* (1984); CAREY MCINTOSH, *The Choice of Life: Samuel Johnson and the World of Fiction* (1973); EDWARD ALAN BLOOM, *Samuel Johnson in Grub Street* (1957); BENJAMIN B. HOOVER, *Samuel Johnson's Parliamentary Reporting* (1953); W.K. WIMSATT, JR., *The Prose Style of Samuel Johnson* (1941, reissued 1972); WILLIAM EDINGER, *Samuel Johnson and Poetic Style* (1977); THOMAS M. CURLEY, *Samuel Johnson and the Age of Travel* (1976); ROBERT DEMARIA, JR., *Johnson's Dictionary and the Language of Learning* (1986); ALLEN REDDICK, *The Making of Johnson's Dictionary, 1746–1773* (1990); and DAISUKE NAGASHIMA, *Johnson the Philologist* (1988).

Books on Johnson's criticism include JEAN H. HAGSTRUM, *Samuel Johnson's Literary Criticism* (1952, reissued 1967); LEOPOLD DAMROSCH, JR., *The Uses of Johnson's Criticism* (1976); CHARLES HINNANT, *"Steel for the Mind": Samuel Johnson and Critical Discourse* (1994); ARTHUR SHERBO, *Samuel Johnson, Editor of Shakespeare* (1956, reprinted 1978); R.D. STOCK, *Samuel Johnson and Neoclassical Dramatic Theory: The Intellectual Context of the Preface to Shakespeare* (1973); G.F. PARKER, *Johnson's Shakespeare* (1989); and EDWARD TOMARKEN, *Samuel Johnson on Shakespeare: The Discipline of Criticism* (1991). Johnson's significant contribution to the law lectures of Robert Chambers is assessed in ROBERT CHAMBERS and SAMUEL JOHNSON, *A Course of Lectures on the English Law*, ed. by THOMAS M. CURLEY, 2 vol. (1986). Johnson's relation to the visual arts is considered in PAUL ALKON and ROBERT FOLKENFLIK, *Samuel Johnson: Pictures and Words* (1984); MORRIS R. BROWNELL, *Samuel Johnson's Attitude to the Arts* (1989); and HERMAN W. LIEBERT, "Portraits of the Author: Lifetime Likenesses of Samuel Johnson," in J. DOUGLAS STEWART and HERMAN W. LIEBERT, *English Portraits of the Seventeenth and Eighteenth Centuries* (1974), pp. 47–88. The fullest exhibition catalogue is KAI KIN YUNG *et al.*, *Samuel Johnson* (1984).

Original collections of essays include ISOBEL GRUNDY (ed.), *Samuel Johnson: New Critical Essays* (1984); PAUL J. KORSHIN (ed.), *Johnson After Two Hundred Years* (1986); JAMES ENGELL (ed.), *Johnson and His Age* (1984); and PREM NATH (ed.), *Fresh Reflections on Samuel Johnson* (1987). (R.Fo.)

Jordan

The Hashemite Kingdom of Jordan (al-Mamlakah al-Urdunnīyah al-Hāshimīyah), an Arab state of Southwest Asia, is a young nation that occupies an ancient land associated with the civilizations of antiquity. It is bounded to the north by Syria, to the east by Iraq, to the southeast and south by Saudi Arabia, and to the west by Israel and the West Bank. (In 1988 Jordan renounced its claims to the West Bank, which had been under its rule from 1948 to 1967.) Jordan has 12 miles (19 kilometres) of coastline on the Gulf of Aqaba in the southwest, where al-ʿAqabah, its only port, is located. The total area is 34,342 square miles (88,946 square kilometres). Jordan's capital and largest city is Amman.

Jordan came into being after World War I, gaining its independence from the United Kingdom as a hereditary constitutional monarchy in 1946. The present ruler, King Hussein ibn Talal, ascended the throne in 1953. He has maintained Jordan's traditional policy of friendship with the West despite strong local and international anti-Western pressures and a number of regional crises.

A country poor in resources, Jordan is surrounded by wealthier and more powerful states. More than 60 percent of its population are Palestinians, the majority of whom arrived as displaced persons following the Arab-Israeli wars of 1948 and 1967 and the Persian Gulf War of 1991. Their presence has added to the economic and political problems of the kingdom.

Jordan has had to depend on outside economic assistance for most of its history. Before 1948 this came primarily from the United Kingdom. Following the Six-Day War with Israel in 1967 aid was provided by the United States and other Western countries (although suspended during the time of the Persian Gulf War), as well as by Arab countries, including Saudi Arabia, Kuwait, and the United Arab Emirates.

This article is divided into the following sections:

Physical and human geography

THE LAND

Relief and drainage. Jordan has three major physiographic regions (from east to west): the desert, the uplands east of the Jordan River, and the Jordan Valley (a branch of the great East African Rift System).

The desert is mostly part of the Syrian (or North Arabian) Desert and is located in the eastern and southern parts of the country, occupying more than four-fifths of its territory. The desert's northern part is composed of volcanic lava and basalt and its southern part of outcrops of sandstone and granite. It is much eroded, primarily by wind. The uplands east of the Jordan River, an escarpment overlooking the rift valley, has an average altitude between 2,000 and 3,000 feet (600 and 900 metres); the elevation increases to about 5,755 feet (1,754 metres) at Mount Ramm in the south. There are outcrops of sandstone, chalk, limestone, and flint extending to the extreme south, where igneous rocks solidified from the molten state predominate. In the northern uplands several valleys and perennial streams run west; around al-Karak they run west, east, and north; south of al-Karak nonperennial valley streams run east toward al-Jafr Depression.

Jordan Valley

The Jordan Valley, some 1,312 feet below sea level at the Dead Sea, contains the lowest point on the Earth's surface. The Jordan River, approximately 186 miles (300 kilometres) in length, meanders south, draining the waters of the Sea of Galilee (Lake Tiberias, or Lake Kinneret [Hebrew name]), the al-Yarmūk River, and the valley streams of both plateaus into the Dead Sea. The soil of its lower reaches is very saline, and the shores of the Dead Sea consist of salt marshes that do not support vegetation. The Dead Sea occupies the central area of the valley. To its south, Wadi al-ʿArabah (also called Wadi al-Jayb), a completely desolate region, is thought to contain mineral resources.

Climate. The climate varies from the Mediterranean type in the west to the desert type in the east and south, but the land is generally arid. The proximity of the Mediterranean Sea is the major climatic influence, although this influence is modified by continental air masses and by altitude. Average monthly temperatures at the capital in the north range between 46° and 78° F (8° and 26° C), while at al-ʿAqabah in the far south they range between 60° and 91° F (16° and 33° C). The prevailing winds throughout the country are westerly to southwesterly, but spells of hot, dry, dusty winds blowing from the southeast off the Arabian Peninsula frequently occur. Known locally as the khamsin, these winds bring the country its most uncomfortable weather. They blow most often in the early and late summer and can last for several days at a time before terminating abruptly as the wind direction changes and much cooler air follows. Rainfall occurs in the short, cool winters, decreasing from 16 inches (400 millimetres) annually in the northwest near the Jordan River to less than 4 inches in the south. The average rainfall in the uplands east of the Jordan River totals about 14 inches annually. The valley itself has a yearly average of 8 inches, and the desert regions receive less than 2 inches. Occasional snow and frost occur in the uplands but are rare in the rift valley. As the population increases, water shortages in the major towns are becoming one of Jordan's crucial problems.

Plant and animal life. The plant and animal life of

Jordan falls into three distinct types: that associated with the Mediterranean, with the steppe (treeless plains), and with the desert. In the uplands the Mediterranean type of scrubby, dense bushes and small trees predominates, while in the drier steppe region to the east sagebrush is most frequent. Grassland is the prevalent vegetation on the steppe, but isolated trees and shrubs, such as lotus fruit and the Mount Atlas pistachio, also occur. In the desert, vegetation grows meagrely in depressions and on the sides and floors of the valleys after the scant winter rains.

There is a variety of wildlife, including wild boars, as well as the ibex, a species of wild goat found in the gorges and in the ʿAyn al-Azraq oasis. Hares, jackals, foxes, wildcats, hyenas, wolves, gazelles, mole rats, mongooses, and a few panthers also inhabit the area. Centipedes, scorpions, and various types of lizards are found as well. Birds include the golden eagle and the vulture, while wild fowl include the pigeon and the partridge.

Settlement patterns. The landscape falls into two regions—the desert zone and the cultivated zone—each of which is associated with its own mode of living. The tent-dwelling nomads (Bedouin, or Badu), who make up 6 to 8 percent of the population, generally inhabit the desert and some areas of the steppe and the uplands. The eastern Bedouin are principally camel breeders and herders, while the western Bedouin are sheep and goat herders. There are some seminomads, in whose existence the modes of life of the desert and the cultivated zones merge. These people adopt a nomadic existence during part of the year but return to their lands and homes in time to practice agriculture. The two largest nomadic tribes of Jordan are Banū Sakhr and al-Ḥuwayṭāt. The grazing grounds of both are entirely within Jordan, as is the case with the smaller tribe of as-Sirḥān. Other, lesser tribes include Banū Hasan, al-Banū Khalid, al-Ajarmeh, al-Adwan, Banū Attiyeh, al-Hajayah, and as-Sleet, as well as the smaller tribes of al-Hawazim, as-Sulaylat, and ash-Sherarat, which traditionally were obliged to pay protection money to larger tribes. The Rwalah tribe, which is not indigenous, passes through Jordan in its yearly wandering from Syria to Saudi Arabia. The tent-dwelling Bedouin population has decreased in number because of successful government efforts at sedentarization; urban Bedouins make up 30 to 40 percent of East Bank Jordanians. Although their political influence has diminished, these martial desert people form the core of Jordan's army, occupying key positions in the military, and they remain committed to King Hussein and the Hashemite regime.

Nomads

Rural residents, including small numbers of nomads, represent about one-fourth of the population. The average village is a cluster of houses and other buildings, including an elementary school and a mosque, with pasturage on the outskirts. A medical dispensary and a post office may be found in the larger villages, together with a general store and a small café, whose owners are usually part-time farmers. Kinship relationships are patriarchal, while extended-family ties govern social relationships and tribal organization.

Of the total population, nearly three-fourths live in urban areas. The main population centres are Amman, az-Zarqāʾ, Irbid, and aṣ-Ṣalt. Many of the smaller towns have only a few thousand inhabitants. Most towns have hospitals, banks, government and private schools, mosques, churches, libraries, and entertainment facilities, and some have institutions of higher learning and newspapers. Amman and az-Zarqāʾ, and to some extent Irbid, have urban characteristics, while smaller towns are more reluctant to accept modernizing influences.

THE PEOPLE

Languages and dialects

Ethnic and religious groups. Nearly all the people are Arabic-speaking, and there are various dialects with local inflections and accents. In addition, there is a significant difference between the written, or classic, Arabic and the colloquial form. The Qaysī-Yemeni dichotomy—a pre-Islāmic split that was introduced to the area with the Arab conquests and that cut across religious and ecological lines—was once an important broad social division. The Arabs, whether Muslim or Christian, used to trace their ancestry from the northern Arabian Qaysī (Maʿdī, Nizārī, Adnanī, or Ismāʿīlī) tribes or from the southern Arabian Yemeni (Banū Kalb or Qahtani) tribes. Only a few tribes and towns have continued to be aware of the split.

The vast majority of the population (more than 90 percent) are Sunnite Muslims; Christians constitute most of the rest. Among Christians, two-thirds adhere to the Rūm, or Greek Orthodox church. Other Christian groups include the Greek Catholics, also called the Melchites, or Catholics of the Byzantine rite, who recognize the supremacy of the pope; the Roman Catholic community, headed by a patriarch appointed by the pope; and the small Syrian Orthodox, or Jacobite, church, whose members use Syriac in their liturgy. Most non-Arab Christians are Armenians; the majority belong to the Gregorian, or Armenian, Orthodox church, the rest to the Armenian Catholic church. There are several Protestant denominations representing relatively recently formed communities whose converts came almost entirely from other Christian sects.

The Druze, an offshoot of the Ismāʿīlī Shīʿite sect, number a few hundred and reside in and around Amman. The Bahāʾī—who in the 19th century also split off from Shīʿite Islām and who number about 1,000—live in al-ʿAdasīyah in the Jordan Valley. The Armenians, Druze, and Bahāʾī are at once religious and ethnic communities. The Shishan (Chechen) are a Circassian Shīʿite Muslim group, numbering about 1,000, who are descended from 19th-century immigrants from the Caucasus Mountains. With the Cherkess (Circassians), who are Sunnite, they make up the most important non-Arab minority. Another small non-Arab group consists of some Turkmen.

Demography. The population structure is predominantly young; persons under the age of 15 constitute the largest component of the population. Only a small percentage of the people are over 65 years of age. The birth rate is high relative to the death rate, producing a natural rate of increase that is about double the world average. Internal migration from rural to urban centres has added a burden to the economy; however, a large number of Jordanians live and work abroad.

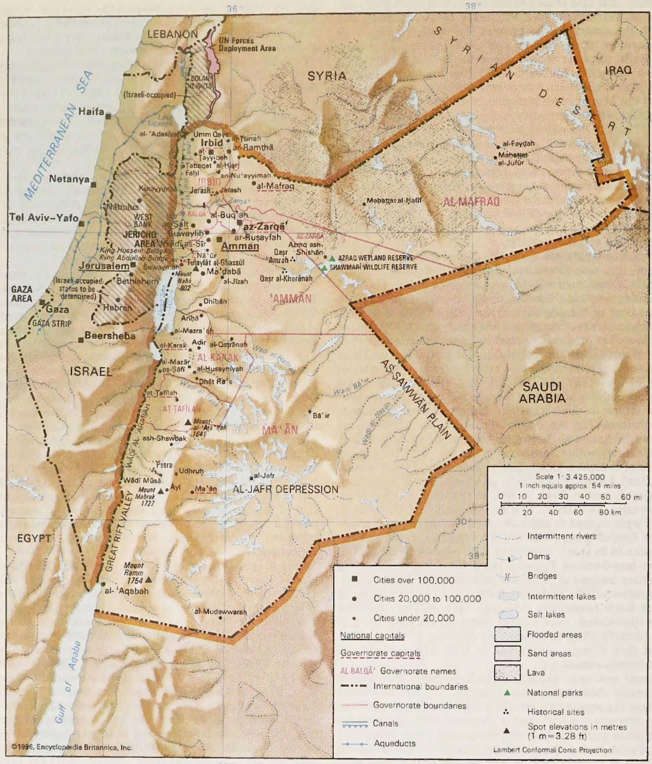
More than two-thirds of Jordan's population are Palestinians, and the differentiation between Palestinians and Transjordanians (those who lived east of the Jordan River before 1948) is the most important distinction in Jordan. The influx of Palestinian refugees has not only altered Jordan's demographic map but also affected its political, social, and economic life. Jordan's original population in the late 1940s was between 200,000 and 250,000. After the 1948–49 Arab-Israeli War and the annexation of the West Bank, Jordanian citizenship was granted to some 400,000 Palestinians who were residents of and remained in the West Bank and to about half a million refugees from the new Israeli state. Many of these refugees settled east of the Jordan River. From 1949 to 1967 Palestinians continued to move east in large numbers. After the 1967 war an estimated 310,000 to 350,000 Palestinians, mostly from the West Bank, sought refuge in Jordan; thereafter immigration from the West Bank continued at a reduced rate. During the 1990–91 Persian Gulf War, an additional 200,000 to 300,000 Palestinians fled Kuwait (or were expelled) to Jordan.

Refugees

Most Palestinians are employed and hold full Jordanian citizenship. By the mid-1990s approximately 1.3 million Palestinians, representing about one-third of Jordan's population, were registered with the United Nations Relief and Works Agency for Palestine Refugees in the Near East (UNRWA). About one-fifth of these refugees were located in camps in Jordan.

THE ECONOMY

Despite its basic problems, the Jordanian economy before 1967 showed resilience and growth. The West Bank, prior to its occupation by Israel in 1967, contributed about one-third of Jordan's total domestic income. Economic growth, halted by war in the second half of 1967, continued thereafter at a slower pace but was revitalized by the implementation of a series of state economic plans. During the Iran-Iraq War (1980–90), trade increased between Jordan and Iraq because of the access Iraq gained thereby



MAP INDEX

Political subdivisions

'Ammān .... 31 40 N 36 30 E
Balqā', al- .... 31 50 N 35 40 E
Irbid .... 32 25 N 35 50 E
Karak, al- .... 31 10 N 35 45 E
Ma'ān .... 30 10 N 36 20 E
Mafraq, al- .... 32 10 N 37 45 E
Ṭafilah, aṭ- .... 30 45 N 35 40 E
Zarqā', az- .... 32 05 N 36 10 E

Cities and towns

'Adasiyah, al- ..... 32 40 N 35 37 E
Adir .... 31 12 N 35 46 E
Amman
('Ammān) ...... 31 57 N 35 56 E
'Aqabah, al- ...... 29 31 N 35 00 E
Ariḥā .... 31 25 N 35 47 E
Ayl .... 30 13 N 35 32 E
Azraq
ash-Shishān ..... 31 50 N 36 49 E
Bā'ir .... 30 46 N 36 41 E
Buq'ah, al- .... 32 05 N 35 53 E
Dhāt Ra's .... 31 00 N 35 46 E
Dhībān .... 31 30 N 35 47 E
Fayḍah, al- .... 32 35 N 38 13 E
Ḥiṣn, al- .... 32 29 N 35 53 E
Ḥusayniyah, al- .... 31 02 N 35 43 E
Irbid .... 32 33 N 35 51 E
Jafr, al- .... 30 18 N 36 13 E
Jarash .... 32 17 N 35 54 E
Jīzah, al- .... 31 42 N 35 57 E
Karak, al- .... 31 11 N 35 42 E
Karāmah, al- ..... 31 57 N 35 35 E
Kurayyimah ...... 32 16 N 35 36 E
Ma'ān .... 30 12 N 35 44 E
Ma'dabā .... 31 43 N 35 48 E
Mafraq, al- .... 32 21 N 36 12 E
Maḥaṭṭat al-Ḥafīf ... 32 12 N 37 08 E
Maḥaṭṭat al-Jufūr ... 32 30 N 38 12 E
Mazār, al- .... 31 04 N 35 42 E
Mazra'ah, al- ..... 31 16 N 35 31 E
Mudawwarah, al- .. 29 19 N 35 59 E
Nā'ūr .... 31 53 N 35 50 E
Nu'ayyimah, an- ... 32 25 N 35 55 E
Qaṭrānah, al- ..... 31 15 N 36 03 E
Ramthā, ar- ...... 32 34 N 36 00 E
Ruṣayfah, ar- ..... 32 01 N 36 03 E
Ṣāfī, aṣ- .... 31 02 N 35 28 E
Ṣalt, aṣ- .... 32 03 N 35 44 E
Shawbak, ash- .... 30 32 N 35 34 E
Ṣuwaylīḥ .... 32 02 N 35 50 E
Suwaymah .... 31 47 N 35 36 E
Ṭabaqat Faḥl ..... 32 27 N 35 37 E
Ṭafilah, aṭ- .... 30 50 N 35 36 E
Ṭayyibah, aṭ- ..... 32 33 N 35 43 E
Ṭurrah, aṭ- .... 32 38 N 35 59 E
Udhruḥ .... 30 20 N 35 36 E
Umm Qays ...... 32 39 N 35 41 E
Wādī as-Sīr ...... 31 57 N 35 49 E
Wādī Mūsā ...... 30 19 N 35 29 E
Zarqā', az- .... 32 05 N 36 06 E

Physical features and points of interest

Allenby Bridge,
see King
Hussein Bridge
Aqaba, Gulf of
(Khalīj
al-'Aqaba) ...... 29 00 N 34 40 E
'Arabah, Wadi al-,
valley .... 30 10 N 35 10 E
'Aṭā'iṭah, Mount
al- .... 30 40 N 35 39 E
Azraq Wetland
Reserve .... 31 51 N 36 51 E
Bādiyat ash-Shām,
see Syrian
Desert
Bā'ir, Wadi .... 31 12 N 37 31 E
Batrā', al-,
see Petra
Dead Sea (al-Baḥr
al-Mayyit) .... 31 30 N 35 30 E
East Ghor Canal
(Qanāt al-Ghawr
ash-Sharqīyah) ... 32 41 N 35 38 E
Great Rift Valley ... 30 00 N 35 00 E
Ḥafīrah, Wadi al- ... 31 15 N 36 02 E
Ḥasā, Wadi al- .... 30 49 N 35 48 E
Ḥaṣāh, Wadi al- ... 30 56 N 37 35 E
Jafr Depression,
al- .... 30 17 N 36 20 E
Jarash, *historical*
site .... 32 17 N 35 54 E
Jordan River
(Nahr al-Urdunn) .. 31 46 N 35 33 E
King Abdullah
Bridge .... 31 48 N 35 33 E
King Hussein
Bridge (Allenby
Bridge) .... 31 52 N 35 32 E
Mabrak, Mount .... 30 13 N 35 29 E
Mayyit,
al-Baḥr al-, see
Dead Sea

Mediterranean Sea . . . . 32 30 N 34 40 E	Tulaylāt al-Ghassūl, *historical site* . . . . 31 48 N 35 36 E
Nabā, Mount . . . . . 31 46 N 35 45 E	Urdunn, Nahr al-, see Jordan River
Petra (al-Batrā'), *historical site* . . . . 30 20 N 35 26 E	Yarmūk, al-, *river* . . 32 38 N 35 34 E
Qaṣr al-Kharānah, *historical site* . . . . 31 44 N 36 28 E	Zarqā, az-, *river* . . . 32 07 N 35 33 E
Qaṣr 'Amrah, *historical site* . . . . 31 48 N 36 35 E	
Ramm, Mount . . . . 29 35 N 35 24 E	
Ṣawwān Plain, aṣ- . 30 45 N 37 15 E	
Shawmari Wildlife Reserve . . . . 31 46 N 36 48 E	
Syrian Desert (Bādiyat ash-Shām) . . . . . . 32 00 N 40 00 E	

to Jordan's port of al-'Aqabah. With Iraq's occupation of Kuwait during the 1990–91 Persian Gulf War, Jordan initially supported Iraqi president Saddam Hussein but eventually agreed to the United Nations' trade sanctions against Iraq, its principal trading partner, and thereby put its whole economy in jeopardy. External emergency aid helped Jordan weather the crisis, and the sudden influx of 200,000–300,000 Palestinians expelled by Kuwait in 1991, many of whom brought capital, helped boost the Jordanian economy.

Sources of revenue

The major sources of revenue are mining and industry, communications and transportation, trade, agriculture, and construction. Income from tourism, which has grown dramatically, is mostly in foreign reserves, and tourism has become a major factor in Jordan's efforts to reduce its balance of payments deficit. Remittances from Jordanians working abroad remain a major source of foreign exchange.

Jordan has lost much of its skilled labour to neighbouring countries—as many as 400,000 people left the kingdom in the early 1980s—although the problem has eased somewhat. This change is a result both of better employment opportunities within Jordan itself and of a curb on foreign labour by the neighbouring Persian Gulf states.

Forest

Resources. Jordan has only some 90,000 acres (36,000 hectares) of forest, most of which are on the rocky highlands. Although unprotected and representing only 1 percent of the country's area, these forests have survived the depredations of villagers and nomads alike, as well as constant overgrazing. The Jordanian government promotes reforestation by providing free seedlings to farmers. In the higher regions of the uplands, the predominant types of trees are the Aleppo oak, the Kermes oak, the Palestinian pistachio, the Aleppo pine, and the Oriental strawberry tree. Wild olives also are found there, and the Phoenician juniper occurs in the regions with lower rainfall.

Less than 10 percent of Jordan is arable, and the country imports foodstuffs to meet its needs. Wheat and barley are the main crops of the rain-fed uplands, and irrigated land in the Jordan Valley produces citrus and other fruits and vegetables (including tomatoes and cucumbers). Pastureland is limited and so degraded that it can barely support Jordan's livestock; it has, moreover, been reduced by the extension of land devoted to olive and fruit trees. Artesian wells have been dug to increase the pasturage area. Sheep and goats are the most important livestock, but there are some cattle, camels, horses, donkeys, and mules. Livestock decreases when droughts occur. There is fishing in the Gulf of Aqaba.

Mineral resources include large deposits of phosphates, potash, limestone, and marble, as well as dolomite, kaolin, and salt. In addition, newly discovered minerals include barite (the principal ore of the metallic element barium), quartzite, gypsum (used as a fertilizer), and feldspar, and there are unexploited deposits of copper, uranium, and shale oil. Manufacturing is concentrated around Amman. The extraction of phosphate, petroleum refining, and cement production are the country's major heavy industries. Articles of food, clothing, and a variety of consumer goods also are produced.

Power in Jordan is generated by fossil fuel, mostly oil. There are several generating plants; the three major power stations, at Amman, al-'Aqabah, and az-Zarqā', are linked by a transmission system. By the late 20th century the government had nearly completed a program to link the major cities and towns by a countrywide grid.

Trade. Exports, though growing, do not cover the value of imports; the deficit is financed by foreign grants, loans, and other forms of capital transfers. Although Jordan's trade deficit has been large, it is offset somewhat by earnings from tourism, remittances sent by Jordanians working abroad, earnings from foreign investments made by the Jordan Central Bank, and subsidies from other Arab and non-Arab governments.

Administration of the economy. The economy is primarily based on private enterprise. The government, in 1962, began planning the economy. All the government plans have given a high priority to agriculture, mining and industry, tourism, and the service industries, in addition to the strengthening and widening of the social, health, welfare, and education infrastructure. Unemployment among the male population remains high, while income per capita has increased dramatically.

Aside from a licensing system, a moderate taxation on luxury items, and the establishment of standards and health measures, both internal and international trade are virtually free of restraint. The government also has participated with private enterprise in establishing the largest mining, industrial, and tourist firms in the country. The government also owns a significant share of the largest companies.

Taxation

Fiscal policy has aimed at increasing revenue by raising various tax rates and by reforming the tax system. Measures applied since the 1960s include increases in customs and excise duties and an increase in income taxes. Although the government has made a great effort to reform the income tax, both to increase revenue and to redistribute income, revenue from indirect taxes continues to exceed that from direct taxes.

The law recognizes the existence of labour unions and employer organizations. The trade-union movement is weak, but this is partly compensated for by the government, which has its own special procedures for settling labour disputes. If governmental efforts fail, the union or employers may resort to the judicial process.

The small size of the Jordanian market, the fluctuations in agricultural production because of irregular rainfall, the lack of capital, political instability in the region, and the presence of refugees all combine to make the continuation of outside help a necessity.

The road system

Transportation. Jordan has a main, secondary, and rural road network, most of which is hard-surfaced. This roadway system, maintained by the Ministry of Public Works and Housing, not only links the major cities and towns but also connects the kingdom with neighbouring countries. Within cities, towns, and villages, however, the local authorities are responsible for road upkeep. The Hejaz-Jordan Railway is government-operated and extends from Dar'ā in the north via Amman to Ma'ān in the south. A southern line operated by the Aqaba Railway Corporation runs to the port of al-'Aqabah and connects with the Hejaz-Jordan Railway at Baṭn al-Ghūl. Rail connections also run from Dar'ā north to Damascus in Syria. The Royal Jordanian Airline links Jordan to Arab, African, Asian, American, and European countries. Queen Alia international airport near al-Jīzah, south of Amman, was opened in 1983. Amman and al-'Aqabah have smaller airports. The expansion of the Jordanian economy, especially the export of phosphate and later the regional transit trade with Iraq, led to the development of the port of al-'Aqabah.

ADMINISTRATION AND SOCIAL CONDITIONS

Government. The 1952 constitution is the most recent of a series of legislative instruments that, both before and after independence, moved toward increased executive responsibility. The constitution declares Jordan to be a constitutional hereditary monarchy with a parliamentary form of government. Islām is the official religion of the state, and Jordan is declared to be part of the Arab *ummah* ("nation"). The king wields wide powers over the executive, legislative, and judicial branches. Jordan's central government is headed by a prime minister appointed by the king; the prime minister then chooses his cabinet. According to the constitution, the appointments of both prime minister and cabinet are subject to parliamentary approval. The cabinet coordinates the work of the different departments and establishes general policy.

Under the constitution the membership of the upper house (the Senate) of the bicameral legislature, composed of *al-aʿyān* ("notables"), is appointed by the king for four years. Elections for *nuwwāb* ("deputies") of the lower house (the House of Representatives) are to be held at least every four years, although elections have been frequently suspended. The ninth parliament, elected in 1965, was prorogued several times before being replaced in 1978 by the National Consultative Council, an appointed body with reduced power that debates government programs and activities. The parliament was reconvened, however, in a special session called in January 1984. Since then the parliament has been periodically suspended: from 1988, when Jordan severed its ties with the West Bank, until 1989 and from August until November 1993, when the country held its first multiparty elections since 1956.

Persons 18 years of age and older may vote provided they meet the legal requirements and are not members of the royal family. Political parties were banned before the elections in 1963, however. Between 1971 and 1976 (when it was abolished), the Arab National Union (originally called the Jordanian National Union) was the only political organization allowed. A national charter approved by King Hussein in 1992 legalized political parties as long as they acknowledged the legitimacy of the monarchy. Among the 20 political parties recognized by the government by the time of the elections held on Nov. 8, 1993, were the Jordanian National Alliance, the Islāmic Action Party, the Jordanian People's Democratic Party, the Communist Party of Jordan, the Jordanian Arab Democratic Party, and the Arab Socialist Baʿath Party.

Jordan is divided into administrative *muḥāfaẓāt* (governorates), which in turn are divided into districts and subdistricts, each of which is headed by an official appointed by the minister of the interior. Cities and towns have mayors and elected councils.

Justice. The judiciary is constitutionally independent, though judges are appointed and dismissed by royal *irādah* ("decree") following a decision of the Justices Council. There are three categories of courts. The first category consists of regular courts, including magistrates' courts, courts of first instance, and courts of appeals and cassation in Amman, which hear appeals passed on from lower appeals courts. The constitution also provides for the Diwān Khāṣṣ (Special Council), which interprets the laws and passes on their constitutionality. The second category consists of Sharīʿah Muslim courts and other religious courts for non-Muslims; these exercise jurisdiction over matters of personal status. The third category consists of special courts, such as land, government, property, municipal, tax, and customs courts.

Three categories of courts

The armed forces. The Jordanian armed forces, which include an air force equipped with modern jet aircraft, developed from the Arab Legion, which was originally commanded by British officers. There is also a small navy (coast guard). The king is commander in chief of the armed forces.

Education. Jordan has three types of schools—government schools, private schools, and the UNRWA schools that have been set up for Palestinian refugee children. Schooling consists of six years of elementary, three years of preparatory, and three years of secondary education. The Ministry of Education supervises all schools and establishes the curricula, teachers' qualifications, and state examinations; it also distributes free books to students in government schools and enforces compulsory education to the age of 14. The majority of the students attend government schools. In addition to Khadduri Agricultural Training Institute, there are agricultural secondary schools as well as a number of vocational, labour, and social affairs institutes, a Sharīʿah (Qurʾānic) seminary, and nursing, military, and teachers' colleges. The University of Jordan was established in Amman in 1962; Yarmūk University was established in 1976. A third university, Muʾtah, opened in 1981.

Health and welfare. Most infectious diseases have been brought under control. The number of physicians has grown rapidly. Comprehensive health facilities are operated by the government, but hospitals are found only in major urban centres. A national health insurance program covers medical, dental, and eye care at a modest cost; service is provided free to the indigent.

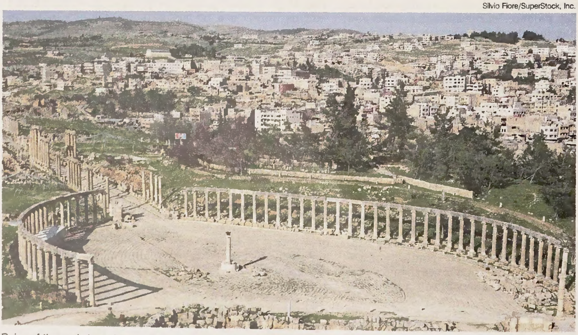
Welfare services were private until the Ministry of Social Affairs was established in 1951. Besides supervising and coordinating social and charitable organizations, the ministry administers welfare programs.

Housing. The housing situation has remained critical despite continuing construction. Surveys conducted in Amman and the eastern Jordan Valley showed that most households live in one-room dwellings. The Housing Corporation and the Jordan Valley Authority build units for low-income families. Urban renewal projects in Amman and az-Zarqāʾ have provided new units and renovated others. The Housing Bank issues home building loans.

CULTURAL LIFE

Culturally, Jordan is an integral part of the Arab world and thus cannot be said to have a separate and distinct culture of its own. As in the rest of the Arab world, the highest form of artistic expression remains oral. Jordan's most famous poet, Muṣṭafā Wahbah aṭ-Ṭāl, ranks among the

Silvio Fiore/SuperStock, Inc.



Ruins of the oval-shaped Forum at Jarash, site of the ancient Roman city of Gerasa; modern Jarash can be seen in the background.

major Arab poets of the 20th century. After World War II a number of important poets and prose writers emerged, though few have achieved an international reputation.

Both private and governmental efforts have been made to foster the arts. Modernity has weakened the traditional Islāmic injunction against the portrayal of animate objects. Thus, in addition to the traditional architecture, decorative design, and various handicrafts, it is possible to find contemporary forms of painting and sculpture.

Folk and popular arts

Folk art survives in tapestry work and in the making of leather, pottery, and ceramics, as well as in the manufacture of wool and goat-hair rugs with varicoloured stripes. Popular culture takes the form of songs, ballads, and storytelling. The villagers have special songs for births, circumcisions, weddings, funerals, planting, plowing, and harvesting. Several types of *debkah* (dances characterized by the pounding of feet on the floor to mark the rhythm) are danced on festive occasions, while the *sahjeh* is a well-known Bedouin dance. The Circassian minority has a sword dance, as well as several other Cossack dances. Governmental interest in preserving folk arts has resulted in the formation of a national troupe that is regularly featured on state radio and television programs.

Newspapers are privately owned but extensively regulated. There are several literary magazines as well as scientific and topical periodicals. Radio and television stations, which are government-owned, feature programs from both Arab and foreign, mostly Western, countries. Most major towns have movie theatres that offer both Arab and foreign films. There is no legitimate theatre in Jordan, but amateur groups perform in institutions of learning, on radio and television, and in the various foreign cultural centres in Amman and Irbid. (K.S.A.J./I.J.B.)

For statistical data on the land and people of Jordan, see the *Britannica World Data* section in the BRITANNICA BOOK OF THE YEAR.

History

Jordan occupies an area rich in archaeological remains and religious traditions. The Jordanian desert was the home of hunters from Lower Paleolithic times; their flint tools have been found widely distributed throughout the region. In the southeastern part of the country, at Mount aṭ-Ṭubayq, there are rock carvings of several prehistoric periods, the earliest of which have been attributed to the Paleolithic-Mesolithic. The site at Tulaylāt al-Ghassūl in the Jordan Valley showing a well-built village with painted plaster walls may represent transitional developments from the Neolithic to the Chalcolithic period.

Early Bronze Age deposits

The Early Bronze Age (*c.* 3000–2100 BCE) is marked by deposits at the base of Dhībān; although many sites have been found in the north of the country, few have been excavated, and little evidence of settlement in this period is found south of ash-Shawbak. The early Bronze Age was terminated by a nomadic invasion that destroyed the principal towns and villages and marked the end of a period of apparently peaceful development. Security was not reestablished again until the advent of the Egyptians after 1580 BCE. It was once thought that the area was unoccupied between 1900 and 1300 BCE, but a systematic archaeological survey has shown that the country had a settled population throughout the period. At Amman a small temple with Egyptian, Mycenaean, and Cypriot imported objects has been found confirming this.

BIBLICAL ASSOCIATIONS

From the Middle Bronze Age onward there are biblical accounts of the area, mentioning kingdoms such as Gilead in the north and those of Moab, in central Jordan, and Midian. At the time of the Exodus, the Israelites tried to pass through Edom in southern Jordan but were refused permission. They were at first repulsed by the Amorites, whom they later defeated. The Israelite tribes of Gad and Reuben and half of the tribe of Manasseh settled in the conquered territory of the Ammonites, Amorites, and Bashan and rebuilt many of the towns they had partially destroyed. A nearly contemporary record of this period is the Mesha or Moabite stone found at Dhībān in 1868 and now in the Louvre Museum, Paris. It is inscribed in an eastern form of Canaanite, closely akin to Hebrew.

The next few centuries (1300–1000 BCE) were marked by constant raiding from both sides of the Jordan. David attacked and devastated Moab and Edom. Although Ammon with its capital, Rabbath Ammon (modern Amman), was held for a time, it regained its independence on the death of David (*c.* 960 BCE). Solomon had a port on the Gulf of Aqaba at Ezion-geber, later Elat in Israel, where copper ore was smelted from mines in the Wadi al-ʿArabah and trade carried on with the southern Arabian states.

Hostilities remained constant between Judah and Edom. A Hebrew king, Amaziah, even captured Sela (Petra), the capital. The next invaders were the Assyrians, who under Adadnirari III (811 or 810–783 BCE) overran the eastern part of the country as far as Edom. Revolts against Assyrian rule occurred in the 760s and 750s, leading to the retaking of the country by Tiglath-pileser III (reigned 745–727 BCE) in 734–733. He then devastated Israel, sent its people into exile, and divided the country into provinces under Assyrian governors. This policy of direct rule continued until the fall of the Assyrian empire in 612 BCE.

It was not until the Hellenistic rule of the Seleucids and the Ptolemies that the country prospered, trade increased, and new towns were built. Rabbath Ammon was renamed Philadelphia, and Jarash became Antioch-on-the-Chrysorrhoas, or Gerasa. Hostilities between the Seleucids and Ptolemies enabled the Nabataeans to extend their kingdom northward and to increase their prosperity based on the caravan trade with Arabia and Syria. The northern part of Jordan was for a time in Jewish hands, and there were constant struggles between the Jewish Maccabees and the Seleucids. It is to this period that the majority of the Dead Sea Scrolls may be attributed.

During 64–63 BCE the kingdom of Nabataea was conquered by the Romans under Pompey, who restored the Hellenistic cities destroyed by the Jews and set up the Decapolis. The country remained independent but paid imperial taxes.

Roman policy seems to have been to maintain Nabataea as a buffer state against the desert tribes. In 25–24 BCE it served as a starting point for Aelius Gallus' ill-starred expedition in search of Arabia Felix. Nabataea was finally absorbed into the Roman Empire by Trajan in 106 CE as the province of Palaestina Tertia. Under Roman rule Jordan prospered, and many new towns and villages were established. The whole country, except the Decapolis, was made part of the new province called Arabia Petraea, with its capital first at Petra and later at Buṣrā ash-Shām in Syria. After 313 CE Christianity became a recognized religion, and a large number of churches were built.

Jordan under Roman rule

THE LATIN KINGDOM AND MUSLIM DOMINATION

The whole area was devastated in the 6th and 7th centuries CE by the intermittent warfare between Byzantium and Sāsānian Persia. In 627 CE the emperor Heraclius finally defeated the Persians and reestablished order in the area, but Byzantium had been gravely weakened by the long struggle and was unable to face the totally unexpected menace of a new power that had arisen in Arabia. In 636 CE the Muslims, led by the famous "Sword of Islām," Khālid ibn al-Walīd, destroyed a Byzantine army at the Battle of the Yarmūk River and brought the greater part of Syria and Palestine under Muslim rule.

The caliphs of the Umayyad dynasty (660–750 CE) established their capital at Damascus and built splendid hunting lodges and palaces in the Jordanian desert. These can still be seen at sites such as Qaṣr ʿAmrah, al-Kharānah, aṭ-Ṭūbah, and Qaṣr al-Mushattā. Many Roman forts were rebuilt. After the seizure of power by the ʿAbbāsids in 750 CE, the capital was transferred to Baghdad, and Syria, which had been the Umayyad metropolitan province, was severely repressed. Jordan, now distant from the centre of power, became a backwater and slowly reverted to the old Bedouin way of life. With the capture of Jerusalem by the crusaders in 1099 CE, the Latin Kingdom of Jerusalem was extended east of the Jordan, and a principality known as Oultre Jourdain was set up. A capital was established at al-Karak. After the crusaders retreated, the history of

Jordan remained uneventful. In the 16th century it submitted to Ottoman rule and became part of the *vilāyet* (province) of Damascus.

In the 19th century the Ottomans settled Circassian, Caucasian, and other refugees in Jordan to protect their communications with Arabia; in 1908 they completed the Hejaz railway linking Damascus and Medina. (V.E.I.)

TRANSJORDAN, THE HASHEMITE KINGDOM, AND THE PALESTINE WAR

During World War I the Arabs joined the British against the Ottomans. In a revolt of 1916, in which they were assisted by Colonel T.E. Lawrence, the Arabs cut the Hejaz railway. In July 1917 the army of Prince Fayṣal ibn Husayn (of the Hashemite dynasty) captured al-ʿAqabah, and by October 1918 Amman and Damascus were in Allied hands. In 1920 the Conference of San Remo (Italy) created two mandates, allotting the one over Palestine to Great Britain and the one over Syria to France. This act effectively separated the area now covered by Israel and Jordan from that of Syria. In November 1920 Abdullah, Fayṣal's brother, arrived in Maʿān, then part of the Hejaz, with 2,000 armed supporters intent on raising the tribes to attack the French, who had forced Fayṣal to relinquish his newly founded kingdom in Syria. By April 1921, however, the British had prevailed upon Abdullah to take over as ruler of what then became known as Transjordan.

The French and British mandates

Effectively, Turkish rule in Transjordan was simply replaced by British rule. The mandate, confirmed by the League of Nations in July 1922, gave the British virtually a free hand in administering the territory, although in September 1922 it was explicitly excluded from the clauses regarding the establishment of "a Jewish national home" and was closed to Jewish immigration. The British recognized Transjordan's independence under the rule of Emir Abdullah on May 25, 1923, codified in a treaty in 1928 (excluding matters of finance and military and foreign affairs, which remained in the hands of a British "resident"). In April 1928 a constitution was promulgated. Full independence was achieved after World War II by a treaty concluded in London on March 22, 1946, and on May 25 Abdullah proclaimed himself king. A new constitution was promulgated, and in 1949 the name of the state was changed to the Hashemite Kingdom of Jordan.

Throughout the interwar years Abdullah had been dependent on British financial support. He also was assisted by them in the formation of an elite force, the Arab Legion, which was commanded and trained by British officers but staffed with Bedouin troops, to maintain order and secure the allegiance of his Bedouin subjects. On May 15, 1948, the day after the Jewish Agency proclaimed the independent state of Israel and immediately after the British withdrew from their Palestine mandate, Transjordan joined its Arab neighbours in the first Arab-Israeli war. The Arab Legion, commanded by Glubb Pasha (John [later Sir John] Bagot Glubb), as well as Egyptian, Syrian, Lebanese, and Iraqi troops entered Palestine. Abdullah's primary purpose, which he had spelled out in secret discussions with Jewish envoys, was to extend his rule to include the area allotted to the Palestinian Arabs under the United Nations partition resolution of November 1947. Accordingly, he engaged his forces in the area of Palestine popularly known as the West Bank and expelled Jewish forces from East Jerusalem (the Old City). When the Jordan-Israel armistice was signed on April 3, 1949, the West Bank and East Jerusalem—an area of about 2,100 square miles—came under Jordanian rule, and the half-million Transjordanians were joined by almost half a million more Palestinian Arabs. This territory was formally annexed by the kingdom in April 1950. Israel and Britain had tacitly agreed to Abdullah keeping the area, but the Arab countries and most of the world opposed the king's action, and only Britain and Pakistan recognized the annexation. The incorporation of the West Bank, with 400,000 Palestinians, into Jordan, as well as a large refugee population that, on the whole, was hostile to the Hashemite regime, brought with them severe economic and political consequences. On the other hand, Abdullah did gain the Muslim shrines such as the Dome of the Rock in Jerusalem's Old City, which compensated for his father's loss of Mecca and Medina at the hands of Ibn Saʿūd a generation earlier.

Death of Abdullah

Abdullah was assassinated at the al-Aqṣā Mosque in Jerusalem on July 20, 1951, by a young Palestinian frustrated by the king's hostility to Palestinian nationalist aspirations. Abdullah's son, Talal, who succeeded him, was declared unfit to rule by parliament because of mental illness after only one year (in August 1952). Talal abdicated in favour of his eldest son, Hussein ibn-Talal, who was crowned king on his 18th birthday, on May 2, 1953.

JORDAN UNDER KING HUSSEIN

Securing the throne, 1953 to *c.* 1960. The history of Jordan after 1953 was largely shaped by King Hussein's policies to secure his throne and to retain or regain the West Bank for the Hashemite dynasty. Jordan's relationship with Israel in the first decade of the Jewish state's existence, although uneasy, was tolerable, though bloody raids and acts of terrorism carried out by each side added to the tension. The kingdom's involvement in the Palestinian question led as much to a contest with Egypt over Jordan's future as it did to a struggle with Israel. In particular, it repeatedly forced Jordan to walk a tightrope between various Arab nations, the Palestinians, and the West and Israel. Thus, popular demonstrations, especially in the West Bank, and pressure from Egypt prevented Hussein in 1955 from signing the pro-Western mutual defense treaty between Great Britain, Turkey, Iran, and Iraq known as the Baghdad Pact, which he had helped initiate. And in 1956 Hussein—bowing to popular pressure and in a show of support for Egyptian efforts at pan-Arab leadership—dismissed his British advisers, including Commander in Chief Glubb, and abrogated the Anglo-Jordanian treaty of 1946. However, when members of the National Guard, drawn mainly from the West Bank, attempted a coup d'état in April 1957, the king, supported by loyal East Bank Bedouins, acted decisively to curb domestic unrest; he purged the legislature of Palestinian nationalists and extremists, banned political parties, and set up a royal dictatorship.

After Egypt and Syria merged in February 1958, establishing the United Arab Republic (UAR; 1958–61), Hussein was persuaded by his cousin King Fayṣal II to join in a federal union with Iraq. In July 1958, however, Fayṣal and his family were killed in an army coup coordinated by Gamal Abdel Nasser. Hussein, realizing his regime was under threat, turned to Great Britain and the United States for assistance. Washington agreed to provide additional military as well as economic aid. The British government, eager to see the pro-Western Hussein secure in Jordan, stationed British paratroops in the country between July and November 1958. This thwarted a further attempt by anti-Hashemite Palestinians supported by Nasser to overthrow the monarchy. By the early 1960s the United States was providing about $100 million per year, enabling economic development, and, despite a number of assassination attempts, the king's future appeared secure.

The PLO and the June 1967 war. The emergence in the late 1960s of the Palestine Liberation Organization (PLO) and the militant group al-Fatah represented a potential threat to Jordan's sovereignty on the West Bank as well as to Israel. In early 1965 al-Fatah, supported by the radical Baʿath Party government in Syria and encouraged by Egypt, began a series of raids against Israel, generally from Jordan, inflicting serious casualties and damage. Israel responded with raids into the West Bank in an effort to force Jordan to quash these military operations. Relations between Jordan and Syria and Egypt and between the Palestinians and Amman deteriorated. Privately, Hussein had been seeking an understanding with Israel over an approach to the external and internal dangers facing the two countries. In late 1966 the Israeli army made a devastating raid into the West Bank village of as-Samu south of Hebron, destroying many of its houses. Hussein responded by attempting to stop the passage of Syrian-based Palestinian guerrillas through Jordan into Israel, eventually breaking off diplomatic ties with Syria (May 23, 1967). However, as tension mounted between Israel and Egypt and Syria

Al-Fatah

in the spring of 1967, Jordan reversed its position and on May 30 signed a defense pact with Egypt and Syria, placing Jordanian forces under Egyptian command. Despite assurances from Israel that Jordan would not be attacked if it remained neutral, Israeli and Jordanian forces clashed in East Jerusalem, and King Hussein joined Egypt and Syria in the third Arab-Israel war in June 1967.

The June 1967 war was a watershed in the modern history of Jordan. Within 48 hours Israeli forces had overrun the entire territory west of the Jordan River, capturing Bethlehem, Hebron, Jericho, Nāblus, Rām Allāh, Janīn, and the city of Jerusalem. Jordan suffered heavy casualties and lost one-third of its most fertile land, and its already overburdened economy was faced with supporting some 200,000 new refugees. Hussein had regarded entering the war as the lesser of two evils; he believed that, if he had not joined Egypt and Syria, they would have supported the Palestinians in overthrowing his regime. The loss of the West Bank and Jerusalem, devastating as it was, was preferable to the loss of his kingdom.

From 1967 to civil war. Following the June war Hussein faced three major problems: how to recover from the economic losses caused by the war, how to live with Israel's occupation of the West Bank and the annexation of East Jerusalem, and how to preserve the Hashemite throne against a considerably augmented and increasingly hostile Palestinian population. The war reversed the progress made in Jordan's economy prior to June 1967, even though Saudi Arabia, Kuwait, and Libya provided the kingdom with foreign aid. Yet within a short period both the United States and Great Britain resumed economic and military aid. In 1971 arrangements also were made with Israel enabling Jordanian cultivation in the Jordan Valley.

Despite the fact that an Arab summit meeting held in Khartoum in August 1967 passed the "three noes" resolution—no peace with Israel, no recognition of Israel, and no negotiations with Israel—King Hussein resumed his secret negotiations with Israel over the disposition of the West Bank and East Jerusalem. Relations with Israel were thus inseparably linked to the future of the Palestinians. Somewhat unrealistically, Hussein sought the return of all the territory lost to Jordanian rule, but, while willing privately to recognize Israel and to cooperate with it across a wide range of issues, he was not prepared to sign a peace treaty with the Jewish state. The two nations were thus no longer enemies and cooperated against PLO terrorism, but there was little progress toward a lasting peace.

Hussein's relations with the PLO, which under the chairmanship of Yāsir ʿArafāt openly challenged the king's control in East Jordan, reached a crisis in September 1970. The radical Marxist Palestinian group, the Popular Front for the Liberation of Palestine (PFLP), hijacked four international airliners and blew up three of them in Dawson's Field, a deserted airstrip in the Jordanian desert. On September 16 the king declared martial law and called in loyal troops, and civil war (later remembered as Black September) erupted. When 250 Syrian tanks entered northern Jordan in support of the PLO, Hussein was forced not only to call upon military assistance from the United States and Great Britain but also to allow Israeli military overflights to attack the Syrian forces. The Syrian forces were defeated, and a peace agreement, in which Hussein made concessions to the PLO, was signed by King Hussein and ʿArafāt in Cairo on Sept. 27, 1970. By July 1971, however, Hussein had forced the PLO guerrillas out of Jordan.

Black September

From 1973 to the *intifada*. King Hussein chose not to join Egypt and Syria in their surprise attack on Israel in the war of October 1973, although he did make a symbolic gesture by sending tanks to assist Syria in the Golan Heights. In negotiations immediately following the war, Hussein once again demanded the return of the West Bank and East Jerusalem from Israel. He was bitter that Israel—in response to pressure from U.S. Secretary of State Henry Kissinger—proposed a withdrawal of its forces from Israeli-occupied Egyptian territory but made no such overtures to Jordan, the neighbour that had stayed out of the war. Yet by August 1974 discussions were under way with Israel over "disengagement accords" that included the recognition of Jordan as speaking for the Palestinians, regional economic cooperation, and tactical cooperation, especially in relation to the threat posed by Palestinian guerrilla groups. However, on Oct. 28, 1974, 20 leaders of the Arab League at an Arab summit meeting in Rabat, Mor., declared that the Palestinian people, under the leadership of the PLO ("their sole legitimate representative"), had the right to establish a national independent authority in liberated Palestine. On Nov. 4, 1974, Hussein announced that Jordan would exclude the West Bank from Jordan and that a federation between Jordan and a Palestinian state was "totally inconceivable," as such a step would inevitably give the Palestinian population a majority and bring about the loss of his kingdom.

The election of the right-wing Likud bloc with Menachem Begin as Israeli prime minister in May 1977 brought relations between Jordan and Israel to a low ebb. Jordan was faced with Begin's determination to annex and retain all of the West Bank, which Israel now called Judaea and Samaria. Begin greatly accelerated the program of constructing Jewish settlements in the West Bank and Gaza. Although Israel was committed to granting autonomy to the Palestinians and to negotiating the future status of the occupied territories under the terms of the Israeli-Egyptian agreement hammered out at Camp David in 1978, Hussein condemned the agreement. He completely broke off the 15-year secret negotiations with Israel. From late 1977 until 1984 Jordanian contacts with Israel came to a virtual halt. Hussein became increasingly alarmed at the rise in popularity in Israel of the view that Jordan was, in fact, the Palestinian state and that the conflict between Israel and the Palestinians would end only when the artificial entity—Jordan—officially became the Palestinian state. Israel's invasion of Lebanon in 1982 fueled fears in Amman that this was the first step in the process of transferring Palestinians to the East Bank.

In the early 1980s Hussein sought an accommodation with ʿArafāt and the PLO. The king realized that ʿArafāt, following his expulsion from Lebanon and the destruction of his bases, was almost entirely friendless and in need of his support. The two men reached a temporary if somewhat uneasy alliance. In order to strengthen his legitimacy in the eyes of Palestinians, Hussein, in 1984, allowed the Palestine National Council (a virtual parliament of the Palestinians) to meet in Amman, and in February 1985 he signed the "Hussein-ʿArafāt" agreement pledging cooperation with the PLO and coordination of a joint peace initiative. In January 1984 Hussein reconvened parliament for the first time since 1974, appointing seven new West Bank representatives and allowing by-elections to be held in March for eight East Bank vacancies in the 60-member House of Representatives. Women were included in the electorate for the first time.

The Amman agreement

In February 1986 Hussein, frustrated by ʿArafāt's ambiguity regarding the PLO's recognition of Israel and the renunciation of terrorism, which in turn confirmed Israeli intransigence, repudiated the Amman agreement with ʿArafāt and broke off negotiations with the PLO. Although the king was careful not to expel the PLO from Jordan entirely, despite an increase in guerrilla violence in the West Bank, he did order the closure of the PLO offices in Amman in March 1986. In a complete turnaround in the Jordanian policy that had been followed since the Arab Rabat summit of 1974, he declared that he would now be responsible for the economic welfare of the West Bank Palestinians and that the West Bank would be included in the new five-year plan for Jordan to be announced in August. The king also approved an increase in the number of Palestinian seats (to about half) in an enlarged National Assembly. His goal was to create a Jordanian-Palestinian-Israeli administration that would make the West Bank independent of the PLO and enable him to reach a settlement with Israel in which he would regain at least partial sovereignty of the area.

By April 1987 Hussein and Shimon Peres, Israel's foreign minister, agreed to a UN-sponsored conference involving all parties to the conflict to seek a comprehensive peace. The Palestinian representatives would be part of a Jordanian-Palestinian delegation. Although the proposal was endorsed by U.S. President Ronald Reagan, the Israeli prime minister Yitzhak Shamir wanted a conference with only Jordan and resisted American pressure for a comprehensive peace conference. King Hussein scored a diplomatic triumph with the staging of an Arab League summit meeting in Amman in November 1987. During this meeting Arab League members agreed to reestablish diplomatic relations with Egypt. More importantly for the king, the Palestinian issue was not the main topic: the Iran-Iraq War, then in its eighth year, took precedence.

The situation changed dramatically in December 1987 with the outbreak of the *intifada*, a Palestinian uprising on the West Bank. King Hussein quickly realized that the uprising was directed against his rule as well as that of the Israelis. His immediate response was to support the *intifada* publicly and to offer aid to families of victims of Israeli reprisals in an effort to deflect hostility to his regime. But the *intifada* leaders (known as the Unified Command) renounced the king's overtures. ʿArafāt quickly assumed the role of spokesman for the revolt. The *intifada* brought to a halt Jordanian and Israeli plans for an economic path to peace. Hussein canceled the five-year plan for the West Bank.

Renouncing claims to the West Bank. An emergency meeting of the Arab League in June 1988 gave the PLO financial control of support for the Palestinians, thereby virtually acknowledging ʿArafāt as their spokesman. In response Hussein, on July 31, renounced all Jordanian claim to the West Bank, allowing the PLO to assume full responsibility. He dissolved the Jordanian parliament (half of whose members were West Bank representatives), ceased salary payments of 21,000 West Bank civil servants, and ordered that West Bank Palestinian passports be converted to two-year travel documents. When the Palestine National Council recognized the PLO as the sole legal representative of the Palestinian people and proclaimed the independence of a purely notional Palestine on Nov. 15, 1988, Hussein immediately extended recognition to the Palestinian entity.

Recognition of the PLO by the Palestine National Council

In November 1989 Jordan held its first parliamentary elections in 22 years. Opposition groups, particularly the fundamentalist Muslim Brotherhood, gained more seats than the pro-government candidates, and the newly elected prime minister, Mudar Badran, promised to lift the martial law in place since 1967—a promise not fully kept until July 1991.

From the Persian Gulf War to peace with Israel. Iraq's invasion of Kuwait in August 1990 and the subsequent 44-day Persian Gulf War in January–February 1991 forced Hussein to choose between two allies, the United States and Iraq. The king leaned heavily toward the "Arab Patriot" Saddam Hussein, who also received a zealous and vocal groundswell of support from the Jordanian people. In addition, trade with Iraq represented 40 percent of the kingdom's gross domestic product. Kuwait's allies immediately cut off all aid to Jordan, imposed an air and sea blockade, and condemned King Hussein's actions. To make matters worse, 200,000 to 300,000 refugees from Kuwait were expelled or fled (back) to Jordan. However, by the end of 1991 the United States and Israel were again seeking Hussein's support for an American-Israeli peace initiative.

The first multiparty general election since 1956 was scheduled for November 1993. In August the king dissolved the 80-member House of Representatives (the lower house of the bicameral National Assembly) and announced that the election would be conducted on a one-person-one-vote system rather than on the old "slate" system that allowed voters to cast as many votes as there were representatives in their constituency. In the election the number of anti-Zionist Islāmic militants—who made up the Islāmic Action Front (IAF), a coalition of Islāmic groupings and the largest of the 20 political parties—was reduced from 36 to 16, giving the king the support he had sought for his policy.

King Hussein expressed public reservations over a PLO-Israeli accord, the Declaration of Principles on Interim Self-Government Arrangements in the occupied territories signed on Sept. 13, 1993, but he stated his willingness to support the Palestinian people. He was concerned over issues relating to Jordan's economic links with the West Bank and the future status of Palestinians in Jordan. About one year later, on Oct. 26, 1994, Jordan and Israel signed a full peace treaty, in which Hussein was recognized as custodian of Muslim holy sites in East Jerusalem.

In January 1995 Hussein signed accords with the PLO pledging support for Palestinian autonomy and the establishment of a Palestinian state including East Jerusalem. The Palestinians nevertheless remained hostile to the peace treaty with Israel, as did Syria and a large segment of the population led by the IAF. Hussein became increasingly frustrated with what he considered to be the obstructionist policies of the Israeli government but, despite failing health, still played a central role as a mediator between Israel and the PLO.

Shortly before his death in February 1999, the king proclaimed his son, Abdullah, to be his successor, rather than his brother, Hassan, who had been the crown prince. King Abdullah continued to carry out his father's policies and focused on integrating economic reforms, bettering Jordan's relations with its Arab neighbours, and improving the status of women. (I.J.B.)

For later developments in the history of Jordan, see the BRITANNICA BOOK OF THE YEAR.

For coverage of related topics in the *Macropædia* and *Micropædia*, see the *Propædia*, sections 911, 924, 962, 96/11, and 978.

BIBLIOGRAPHY

General works. Overviews of all aspects of the country include HELEN CHAPIN METZ (ed.), *Jordan: A Country Study*, 4th ed. (1991); and PETER GUBSER, *Jordan: Crossroads of Middle Eastern Events* (1983). Further resources may be found in IAN J. SECCOMBE (compiler), *Jordan* (1984), an annotated bibliography; and the bibliography in PETER GUBSER, *Historical Dictionary of the Hashemite Kingdom of Jordan* (1991).

Physical and human geography. COLBERT C. HELD, *Middle East Patterns: Places, Peoples, and Politics*, 2nd ed. (1994), places Jordan in a larger context. GUY MOUNTFORT, *Portrait of a Desert* (1965), gives an illustrated description of life in the desert of Jordan; while RAMI G. KHOURI, *The Jordan Valley: Life and Society Below Sea Level* (1981), addresses agriculture and development on both sides of the river. CHRISTINE OSBORNE, *An Insight and Guide to Jordan* (1981), offers an overview of the landscape, society, and culture. Studies in social anthropology include RICHARD T. ANTOUN, *Arab Village: A Social Structural Study of a Transjordanian Peasant Community* (1972); PETER GUBSER, *Politics and Change in Al-Karak, Jordan;* and PAUL A. JUREIDINI and R.D. MCLAURIN, *Jordan: The Impact of Social Change on the Role of the Tribes* (1984). NORMAN N. LEWIS, *Nomads and Settlers in Syria and Jordan, 1800–1980* (1987), traces the shift from a grazing economy to a sedentary agricultural society. Economic and political conditions are addressed in MICHAEL P. MAZUR, *Economic Growth and Development in Jordan* (1979); BICHARA KHADER and ADNAN BADRAN (eds.), *The Economic Development of Jordan* (1987), a collection of essays; NASEER H. ARURI, *Jordan: A Study in Political Development (1921–1965)* (1972); and RODNEY WILSON (ed.), *Politics and the Economy in Jordan* (1991).

History. The history of Jordan within the region is studied in WILLIAM L. CLEVELAND, *A History of the Modern Middle East* (1994). The most accessible general study of Jordanian history is KAMAL SALIBI, *The Modern History of Jordan* (1993). ADNAN HADIDI (ed.), *Studies in the History and Archaeology of Jordan*, 3 vol. (1982–87), contains detailed analyses of ancient and medieval history. Studies of Jordan include URIEL DANN, *Studies in the History of Transjordan, 1920–1949* (1984). MARY C. WILSON, *King Abdullah, Britain, and the Making of Jordan* (1987); and AVI SHLAIM, *Collusion Across the Jordan: King Abdullah, the Zionist Movement, and the Partition of Palestine* (1988), also available in an abridged ed., *The Politics of Partition* (1990), examine the role of King Abdullah at critical points in Jordan's history. AMNON COHEN, *Political Parties in the West Bank Under the Jordanian Regime, 1949–1967*, trans. from Hebrew (1982); SHAUL MISHAL, *West Bank/East Bank: The Palestinians in Jordan, 1949–1967* (1978); and CLINTON BAILEY, *Jordan's Palestinian Challenge, 1948–1983: A Political History* (1984), discuss Jordan and the Palestinians. P.J. VATIKIOTIS, *Politics and the Military in Jordan*. VALERIE YORKE, *Domestic Politics and Regional Security: Jordan, Syria, and Israel* (1988), examines Jordan's domestic political dynamics. MADIHA RASHID AL MADFAI, *Jordan, the United States, and the Middle East Peace Process, 1974–1991* (1993), explores Jordan's diplomatic role. (K.S.A.J./I.J.B.)

Judaism

Judaism, the religion of the Jews, is the complex expression of a religious and ethnic community, a way of life as well as a set of basic beliefs and values, which is discerned in patterns of action, social order, and culture as well as in religious statements and concepts. The first section of this article treats the history of Judaism in the broadest and most complete sense, from the early ancestral beginnings of the Jewish people down to contemporary times. In the second section the beliefs, practices, and culture of Judaism are discussed. Dates are listed throughout as BCE (Before the Common Era = BC) and CE (Common Era = AD).

The article is organized as follows:

THE HISTORY OF JUDAISM

It is history that provides the clue to an understanding of Judaism, for its primal affirmations appear in early historical narratives. Many contemporary scholars agree that although the biblical (Old Testament) tales report contemporary events and activities, they do so for essentially theological reasons. Such a distinction, however, would have been unacceptable to the authors, for their understanding of events was not superadded to but was contemporaneous with their experience or report of them. For them, it was primarily within history that the divine presence was encountered. God's presence was also experienced within the natural realm, but the more immediate or intimate disclosure occurred in human actions. Although other ancient communities saw a divine presence in history, this was taken up in its most consequent fashion within the ancient Israelite community and has remained, through many developments, the focus of its descendants' religious affirmations. It is this particular claim—to have experienced God's presence in human events—and its subsequent development that is the differentiating factor in Jewish thought. As ancient Israel believed itself through its history to be standing in a unique relationship to the divine, this basic belief affected and fashioned its life-style and mode of existence in a way markedly different from groups starting with a somewhat similar insight. The response of the people Israel to the divine presence in history was seen as crucial not only for itself but for all mankind. Further, God had—as person—in a particular encounter revealed the pattern and structure of communal and individual life to this people. Claiming sovereignty over the people because of his continuing action in history on its behalf, he had established a *berit* ("covenant") with it and had required from it obedience to his Torah (teaching). This obedience was a further means by which the divine presence was made manifest—expressed in concrete human existence. The corporate life of the chosen community was thus a summons to the rest of mankind to recognize God's presence, sovereignty, and purpose—the establishment of peace and well-being in the universe and in mankind.

God's presence in history

History, moreover, disclosed not only God's purpose but also manifested man's inability to live in accord with it. Even the chosen community failed in its obligation and had, time and again, to be summoned back to its responsibility by divinely called spokesmen—the prophets—who warned of retribution within history and argued and reargued the case of affirmative human response. Israel's role in the divine economy and thus Israel's particular culpability were dominant themes sounded against the motif of fulfillment, the ultimate triumph of the divine purpose, and the establishment of divine sovereignty over all mankind.

General observations

NATURE AND CHARACTERISTICS

In nearly 4,000 years of historical development, the Jewish people and their religion have displayed both a remarkable adaptability and continuity. In their encounter with the great civilizations, from ancient Babylonia and Egypt down to Western Christendom and modern secular culture, they have assimilated foreign elements and integrated them into their own socioreligious system, thus maintaining an unbroken line of ethnic and religious tradition. Furthermore, each period of Jewish history has left behind it a specific element of a Judaic heritage that continued to influence subsequent developments, so that the total Jewish heritage at any time is a combination of all these successive elements along with whatever adjustments and accretions are imperative in each new age.

The fundamental teachings of Judaism have often been grouped around the concept of an ethical (or ethical-historical) monotheism. Belief in the one and only God of Israel has been adhered to by professing Jews of all ages and all shades of sectarian opinion. By its very nature monotheism ultimately postulated religious universalism, although it could be combined with a measure of particularism. In the case of ancient Israel (see below *Biblical Judaism*), particularism took the shape of the doctrine of election; that is, of a people chosen by God as "a kingdom of priests and a holy nation" to set an example for all mankind. Such an arrangement presupposed a covenant between God and the people, the terms of which the chosen people had to live up to or be severely punished. As the 8th-century-BCE prophet Amos expressed it: "You only have I known of all the families of the earth; therefore I will punish you for all your iniquities." Further, it was a concept that combined with the messianic idea, according to which, at the advent of the Redeemer, all nations would see the light, give up war and strife, and follow the guidance of the Torah (divine guidance, teaching, or law) emanating from Zion (a hill in Jerusalem that has a special spiritual significance). With all its variations in detail, messianism has, in one form or another, permeated Jewish thinking throughout the ages and, under various guises, has coloured the outlook of many secular-minded Jews (see also DOCTRINES AND DOGMAS: *Eschatology*).

Law became the major instrumentality by which Judaism was to bring about the reign of God on earth. In this case law meant not only what the Romans called *jus* (human law) but also *fas*, the divine or moral law that embraces practically all domains of life. The ideal, therefore, as expressed in the Ten Commandments, was a religioethical conduct that involved ritualistic observance as well as individual and social ethics, a liturgical–ethical way constantly expatiated on by the prophets and priests, rabbinic sages, and philosophers. Such conduct was to be placed in the service of God, as the transcendent and immanent Ruler of the universe, and as such the Creator and propelling force of the natural world, and also as the One giving guidance to history and thus helping man to overcome the potentially destructive and amoral forces of nature. According to Judaic belief, it is through the historical evolution of man, and particularly of the Jewish people, that the divine guidance of history constantly manifests itself and will ultimately culminate in the messianic age. Judaism, whether in its "normative" form or its sectarian deviations, never completely departed from this basic ethical–historical monotheism.

(S.W.B./L.H.S.)

PERIODIZATION

The division of the millennia of Jewish history into periods—a procedure frequently dependent on individual preferences—has not been devoid of theological or scholarly presuppositions. The Christian world long believed that until the rise of Christianity the history of Judaism was but a "preparation for the Gospel" (*preparatio evangelica*) followed by the "manifestation of the Gospel" (*demonstratio evangelica*) as revealed by Christ and the Apostles. This formulation could be theologically reconciled with the assumption that Christianity had been preordained even before the creation of the world.

On the other hand, 19th-century biblical scholars moved the decisive division back into the period of the Babylonian Exile and restoration of the Jews to Judah (6th–5th centuries BCE). They asserted that after the first fall of Jerusalem (586 BCE) the ancient "Israelitic" religion gave way to a new form of the "Jewish" faith, or Judaism, as formulated by Ezra the Scribe and his school (5th century BCE). A German historian, Eduard Meyer, in 1896 published *Die Entstehung des Judentums* ("The Origin of Judaism"), in which he placed the origins of Judaism in the Persian period (see below *Biblical Judaism*) or the days of Ezra and Nehemiah (5th century BCE) and actually attributed to Persian imperialism an important role in shaping the new emergent Judaism.

These theories have been discarded by most scholars, however, in the light of a more comprehensive knowledge of the ancient Middle East and the abandonment of a theory of gradual evolutionary development that was dominant at the beginning of the 20th century. Most Jews share a long-accepted notion that there never was a real break in continuity and that Mosaic-prophetic-priestly Judaism was continued, with but few modifications, in the work of the Pharisaic and rabbinic sages (see below *Rabbinic Judaism*) well into the modern period. Even today the various Jewish groups, whether Orthodox, Conservative, or Reform, all claim direct spiritual descent from the Pharisees and the rabbinic sages. In actual historical development, however, many deviations have occurred from so-called normative or rabbinic Judaism.

In any case, the history of Judaism here is viewed as falling into the following major periods of development: biblical Judaism (*c.* 20th–4th century BCE), Hellenistic Judaism (4th century BCE–2nd century CE), rabbinic Judaism (2nd–18th century CE), and modern Judaism (*c.* 1750 to the present). (S.W.B.)

Biblical Judaism (20th–4th century BCE)

THE ANCIENT MIDDLE EASTERN SETTING

The family of the Hebrew patriarchs (Abraham, Isaac, and Jacob) is depicted in the Bible as having had its chief seat in the northern Mesopotamian town of Harran—then (mid-2nd millennium BCE) belonging to the Hurrian kingdom of Mitanni. From there Abraham, the founder of the Hebrew people, is said to have migrated to Canaan (comprising roughly the region of modern Israel and Lebanon)—throughout the biblical period and later ages a vortex of west Asian, Egyptian, and east Mediterranean ethnoculture. Thence the Hebrew ancestors of the people of Israel (named after the patriarch Jacob, also called Israel) migrated to Egypt, where they lived in servitude, and a few generations later returned to occupy part of Canaan. The Hebrews were seminomadic herdsmen and occasionally farmers, ranging close to towns and living in houses as well as tents.

Early Israelite culture

The initial level of Israelite culture resembled that of its surroundings; it was neither wholly original nor primitive. The tribal structure resembled that of West Semitic steppe dwellers known from the 18th-century-BCE tablets excavated at the north central Mesopotamian city of Mari; their family customs and law have parallels in Old Babylonian and Hurro-Semite law of the early and middle 2nd millennium. The conception of a messenger of God that underlies biblical prophecy was Amorite (West Semitic) and found in the tablets at Mari. Mesopotamian religious and cultural conceptions are reflected in biblical cosmogony, primeval history (including the Flood story in Gen. 6:9–8:22), and law collections. The Canaanite component of Israelite culture consisted of the Hebrew language and a rich literary heritage—whose Ugaritic form (which flourished in the northern Syrian city of Ugarit from the mid-

15th century to about 1200 BCE) illuminates the Bible's poetry, style, mythological allusions, and religiocultic terms. Egypt provides many analogues for Hebrew hymnody and wisdom literature. All the cultures among which the patriarchs lived had cosmic gods who fashioned the world and preserved its order, including justice; all had a developed ethic expressed in law and moral admonitions; and all had sophisticated religious rites and myths.

Though plainer when compared with some of the learned literary creations of Mesopotamia, Canaan, and Egypt, the earliest biblical writings are so imbued with contemporary ancient Middle Eastern elements that the once-held assumption that Israelite religion began on a primitive level must be rejected. Late-born amid high civilizations, the Israelite religion had from the start that admixture of high and low features characteristic of all the known religions of the area. Implanted on the land bridge between Africa and Asia, it was exposed to crosscurrents of foreign thought throughout its history.

THE PRE-MOSAIC PERIOD: THE RELIGION OF THE PATRIARCHS

Israelite tradition identified YHWH (by scholarly convention pronounced Yahweh), the God of Israel, with the Creator of the world, who had been known to and worshipped by men from the beginning of time. Abraham (perhaps 19th or 18th–17th centuries BCE) did not discover this God, but entered into a new covenant relation with him, in which he was promised the land of Canaan and a numerous progeny. God fulfilled that promise through the actions of the 13th-century-BCE Hebrew leader Moses: he liberated the people of Israel from Egypt, imposed Covenant obligations on them at Mt. Sinai, and brought them to the promised land.

The God of the patriarchs

Historical and anthropological studies present formidable objections to the continuity of YHWH worship from Adam (the biblical first man) to Moses, and the Hebrew tradition itself, moreover, does not unanimously support even the more modest claim of the continuity of YHWH worship from Abraham to Moses. Against it is a statement in chapter 6, verse 3, of Exodus that God revealed himself to the patriarchs not as YHWH but as El Shaddai—an epithet (of unknown meaning) the distribution of which in patriarchal narratives and Job and other poetical works confirms its archaic and unspecifically Israelite character. Comparable is the distribution of the epithet El Elyon (God Most High). Neither of these epithets appears in postpatriarchal narratives (excepting the Book of Ruth). Other compounds with El are unique to Genesis: El Olam (God the Everlasting One), El Bethel (God Bethel), and El Ro'i (God of Vision). An additional peculiarity of the patriarchal stories is their use of the phrase "God of my [your, his] father." All of these epithets have been taken as evidence that patriarchal religion differed from the worship of YHWH that began with Moses. A relation to a patron god was defined by revelations starting with Abraham (who never refers to the God of his father) and continuing with a succession of "founders" of his worship. Attached to the founder and his family, as befits the patron of wanderers, this unnamed deity (if indeed he was one only) acquired various Canaanite epithets (El, Elyon, Olam, Bethel, *qone eretz* [possessor of the Land]) only after their immigration into Canaan. Whether the name of YHWH was known to the patriarchs is doubtful. It is significant that while the epithets Shaddai and El occur frequently in pre-Mosaic and Mosaic-age names, YHWH appears as an element only in the names of Yehoshua' (Joshua) and perhaps of Jochebed—persons who were closely associated with Moses.

Important sites and regions of biblical Judaism.

The patriarchs are depicted as objects of God's blessing, protection, and providential care. Their response is loyalty and obedience and observance of a cult whose ordinary expression is sacrifice, vow, and prayer at an altar, stone pillar, or sacred tree. Circumcision was a distinctive mark of the cult community. The eschatology (doctrine of ultimate destiny) of their faith was God's promise of land and a great progeny. Any flagrant contradictions between patriarchal and later mores have presumably been censored; yet distinctive features of the post-Mosaic religion are absent. The God of the patriarchs shows nothing of YHWH's "jealousy"; no religious tension or contrast with their neighbours appears, and idolatry is scarcely an issue. The patriarchal covenant differed from the Mosaic Sinaitic Covenant in that it was modelled upon a royal grant to favourites and contained no obligations, the fulfillment of which was to be the condition of their happiness. Evidently not the same as the later religion of Israel, patriarchal religion prepared the way for it in its familial basis, its personal call by the deity, and its response of loyalty and obedience to him.

Little can be said of the relation of the religion of the patriarchs to the religions of Canaan. Known points of contact between the two are the divine epithets mentioned above. Like the God of the fathers, El, the head of the Ugaritic pantheon was depicted both as a judgmental and a compassionate deity. Baal (Lord), the aggressive young agricultural deity of Ugarit, is remarkably absent from Genesis. Yet the socioeconomic situation of the patriarchs was so different from the urban, mercantile, and monarchical background of the Ugaritic myths as to render any comparisons highly questionable.

THE MOSAIC PERIOD: FOUNDATIONS OF THE ISRAELITE RELIGION

The Egyptian sojourn. According to Hebrew tradition, a famine caused the migration to Egypt of the band of 12 Hebrew families that later made up a tribal league in the land of Israel. The schematic character of this tradition does not impair the historicity of a migration to Egypt, an enslavement by Egyptians, and an escape from Egypt under an inspired leader by some component of the later league of Israelite tribes. To disallow these events would make their centrality as articles of faith in the later religious beliefs of Israel inexplicable.

The importance of Moses, the Exodus, and the Sinai Covenant

Tradition gives the following account of the birth of the nation. At the Exodus from Egypt (13th century BCE), YHWH showed his faithfulness and power by liberating Israel from bondage and punishing their oppressors with plagues and drowning at the sea. At Sinai, he made Israel his people and gave them the terms of his Covenant, regulating their conduct toward him and each other so as to make them a holy nation. After sustaining them miraculously during their 40-year wilderness trek, he enabled them to take the land that he had promised to their fathers, the patriarchs. Central to these events is God's apostle, Moses, who was commissioned to lead Israel out of Egypt, mediate God's Covenant to them, and bring them to Canaan.

Behind the legends and the multiform law collections, a historical figure must be posited to whom the legends and the legislative activity could be attached. And it is precisely Moses' unusual combination of roles that makes him credible as a historical figure. Like Muḥammad at the birth of Islām, Moses fills oracular, legislative, executive, and military functions. The main institutions of Israel are his creation: the priesthood and the sacred shrine, the Covenant and its rules, the administrative apparatus of the tribal league. Significantly, though Moses is compared to a prophet in various texts in the Pentateuch (the first five books of the Bible), he is never designated as one—the term being evidently unsuited for so comprehensive and unique a figure.

Mosaic religion. The distinctive features of Israelite religion appear with Moses. The proper name of Israel's God, YHWH, was revealed and interpreted to Moses as meaning *ehye asher ehye*—an enigmatic phrase (literally meaning "I am/shall be what I am/shall be") of infinite suggestiveness. The Covenant, defining Israel's obligations, is ascribed to Moses' mediation. Although it is impossible to determine what rulings go back to Moses, the Decalogue, or Ten Commandments, presented in chapter 20 of Exodus and chapter 5 of Deuteronomy, and the larger and smaller Covenant codes in Ex. 20:22–23:33; 34:11–26) are held by critics to contain early covenant law. From them, the following features may be noted: (1) The rules are formulated as God's utterances—*i.e.*, expressions of his sovereign will. (2) They are directed toward, and often explicitly addressed to, the people at large; Moses merely conveys the sovereign's message to his subjects. (3) Publication being of the essence of the rules, the people as a whole are held responsible for their observance (see also DOCTRINES AND DOGMAS: *Covenant*).

The liberation from Egypt laid upon Israel the obligation of exclusive loyalty to YHWH. This meant eschewing all other gods—including idols venerated as such—and the elimination of all magical recourses. The worship of YHWH was aniconic (without images); even such figures as might serve in his worship were banned—apparently owing to the theurgic overtones (the implication that through them men may influence or control the god by fixing his presence in a particular place and making him accessible). Though a mythological background lies behind some cultic terminology (*e.g.*, "a pleasing odor to YHWH," "my bread"), sacrifice is rationalized as tribute or (in priestly writings) is regarded purely as a sacrament; *i.e.*, as a material means of relating to God. Hebrew festivals also have no mythological basis; they either celebrate God's bounty (*e.g.*, at the ingathering of the harvest) or his saving acts (*e.g.*, the festival of unleavened bread, which is a memorial of the Exodus).

By courtesy of the Societe des Editions Verve, from "La Bible illustree par Chagall," *Verve*, no. 33/34

Moses breaking the tablets of the Law on Mt. Sinai in anger over the Israelites' worship of the golden calf. Etching by Marc Chagall.

Social values and concerns

The values of life and limb, labour, and social solidarity are protected in the rules on relations between man and man. The involuntary perpetual slavery of Hebrews is abolished, and a seven-year limit is set on bondage. The humanity of slaves is defended: one who beats his slave to death is liable to death; if he maims a slave he must set the slave free. A murderer is denied asylum and may not ransom himself from death, while for deliberate and severe bodily injuries the lex talionis ("an eye for an eye" principle) is ordained. Harm to property or theft is punished monetarily, never by death.

Moral exhortations call for solidarity with the poor and the helpless, for brotherly assistance to fellows in need. Institutions are created (*e.g.*, the sabbatical, or seventh, fallow year, in which land is not cultivated) to embody them in practice.

Since the goal of the people was the conquest of a land, their religion had warlike features. Organized as an army (called "the hosts of YHWH" in Ex. 12:41), they encamped in a protective square around their palladium—the tent housing the ark in which the stone "tablets of the Covenant" rested. When journeying, the sacred objects were carried and guarded by the Levites (a tribe serving religious functions), whose rivals, the Aaronites, had a monopoly on the priesthood. God, sometimes called "the warrior," marched with the army; in war, part of the booty was delivered to his ministers.

THE PERIOD OF THE CONQUEST AND SETTLEMENT OF CANAAN

The conquest of Canaan was remembered as a continuation of God's marvels at the Exodus. The Jordan River was split asunder, Jericho's walls fell at Israel's shout; the enemy was seized with divinely inspired terror; the sun stood still in order to enable Israel to exploit its victory. Such stories are not necessarily the work of a later age; they reflect rather the impact of these victories on the actors in the drama, who felt themselves successful by the grace of God.

A complex process of occupation, involving both battles of annihilation and treaty arrangements with the natives, has been simplified in the biblical account of Joshua's wars. Gradually, the unity of the invaders dissolved (most scholars believe that the invading element was only part of the Hebrew settlement in Canaan; other Hebrews, long since settled in Canaan from patriarchal times, then joined the invaders' covenant league). Individual tribes made their way with more or less success against the residue of Canaanite resistance. New enemies, Israel's neighbours to the east and west, appeared, and the period of the judges (leaders, or champions) began.

The religious situation during the time of the Book of Judges

The Book of Judges, the main witness for the period, does not speak with one voice on the religious situation. Its editorial framework describes repeated cycles of apostasy, oppression, appeal to God, and relief through a Godsent champion. The premonarchic troubles (before the kingship of Saul; see below) caused by the weakness of the disunited tribes were thus accounted for by the covenantal sin of apostasy. The individual stories, however, present a different picture. Apostasy does not figure in the exploits of the judges Ehud, Deborah, Jephthah, and Samson; YHWH has no rival, and faith in him is periodically confirmed by the saviours he sends to rescue Israel from their neighbours.

This faith is shared by all the tribes; and it is owing to their common cult that a Levite from Bethlehem could serve first at an Ephraimite and later also at a Danite sanctuary. The religious bond, preserved by the common cult, was enough to enable the tribes to act more or less in concert under the leadership of elders or an inspired champion in time of danger or religious scandal.

To be sure, both written and archaeological testimonies point to the Hebrews' adoption of Canaanite cults—the Baal worship of Gideon's family and neighbours in Ophrah in Judges, chapter 6, is an example. The many cultic figurines (usually female) found in Israelite levels of Palestinian archaeological sites also give colour to the sweeping indictments of the framework of the Book of Judges. But these phenomena belonged to the private, popular religion; the national God, YHWH, remained one—Baal sent no prophets to Israel—though YHWH's claim to exclusive worship was obviously not effectual. Nor did his cult conform with later orthodoxy; Micah's idol in Judges, chapter 17, and Gideon's ephod (priestly or religious garment) were considered apostasies by the editor, in accord with the dogma that other than orthodoxy there is only apostasy—heterodoxy (nonconformity) being unrecognized and simply equated with apostasy.

To the earliest sanctuaries and altars honoured as patriarchal foundations—at Shechem, Bethel, Beersheba, and Hebron in Cisjordan (west of the Jordan); at Mahanaim, Penuel, and Mizpah in Transjordan (east of the Jordan)—were now added new ones at Dan, Shiloh, Ramah, Gibeon, and Gilgal, among others. A single priestly family could not operate all these establishments, and Levites rose to the priesthood; at private sanctuaries even non-Levites might be consecrated as priests. The ark of the Covenant was housed in the Shiloh sanctuary, staffed by priests of the house of Eli, who traced their consecration back to Egypt. But the ark remained a portable palladium in wartime; Shiloh was not regarded as its final resting place. The law in Exodus, chapter 20, verses 24–26, authorizing a plurality of altar sites and the simplest forms of construction (earth and rough stone) suited the plain conditions of this period.

THE PERIOD OF THE UNITED MONARCHY

The religiopolitical problem. The loose, decentralized tribal league could not cope with the constant pressure of external enemies—camel-riding desert marauders who pillaged harvests annually or iron-weaponed Philistines (an Aegean people settling coastal Palestine *c.* 12th century BCE) who controlled key points in the hill country occupied by Israelites. In the face of such threats to the Israelites, local, sporadic, God-inspired saviours had to be replaced by a continuous central leadership that could mobilize the forces of the entire league and create a standing army. Two attitudes were distilled in the crisis, one conservative and antimonarchic, the other progressive and promonarchic. The conservative appears first in Gideon's refusal, in Judges, chapter 8, verse 23, to found a dynasty: "I will not rule you," he tells the people, "my son will not rule over you; YHWH will . . . !" This theocratic view pervades one of the two contrasting accounts of the founding of the monarchy fused in chapters 8–12 of the First Book of Samuel: the popular demand for a king was viewed as a rejection of the kingship of God, which was embodied in a series of inspired saviours from Moses and Aaron, through Jerubbaal, Bedan, and Jephthah, to Samuel. The other account depicts the monarchy as a gift of God, designed to rescue his people from the Philistines (I Sam. 9:16). Both accounts represent the seer-judge Samuel as the key figure in the founding of Israel's monarchy, and it is not unlikely that the two attitudes struggled in his breast.

Conflicting views regarding the monarchy

By courtesy of the Bollingen Foundation, photograph, Fred Anderegg

The triumph of the ark of the Covenant over paganism, a representation of the antagonism between Judaism and Hellenistic paganism. It was inspired by the biblical story (I Sam. 5:1–5) of the ark, which was captured in battle by the Philistines and which toppled the cult images of Dagon. Mural painting from the synagogue at Doura-Europus, Syria, 3rd century CE.

The Benjaminite Saul was made king (*c.* 1020 BCE) by divine election and by popular acclamation after his victory over the Ammonites (a Transjordanian Semitic people), but his career was clouded by conflict with Samuel, the major representative of the old order. Saul's kingship was bestowed by Samuel and had to be accommodated to the ongoing authority of that man of God. The two accounts of Saul's rejection by God (through Samuel) involve his usurpation of the prophet's authority. King David, whose forcefulness and religiopolitical genius established the monarchy (*c.* 1000 BCE) on an independent spiritual footing, resolved the conflict.

The Davidic monarchy. The essence of the Davidic innovation was the idea that, in addition to divine election through Samuel and public acclamation, David had God's promise of an eternal dynasty (a conditional, perhaps earlier, and an unconditional, perhaps later, form of this promise exist in Psalms, 132 and II Samuel, chapter 7, respectively). In its developed form, the promise was conceived of as a covenant with David, parallelling the Covenant with Israel and instrumental in the latter's fulfillment; *i.e.*, that God would channel his benefactions to Israel through the chosen dynasty of David. With this new status came the inviolability of the person of God's anointed (a characteristically Davidic idea) and a court rhetoric—adapted from pagan models—in which the king was styled "the [firstborn] son of God." An index of the king's sanctity was his occasional performance of priestly duties. Yet the king's mortality was never forgotten—he was never deified; prayers and hymns might be said on his behalf, but they were never addressed to him as a god.

David captured the Jebusite stronghold of Jerusalem and made it the seat of a national monarchy (Saul had never moved the seat of his government from his home town, the Benjaminite town of Gibeah, about three miles north of Jerusalem). Then, fetching the ark from an obscure retreat, David installed it in his capital, asserting his royal prerogative (and obligation) to build a shrine for the national God—at the same time joining the symbols of the dynastic and the national covenants. This move of political genius linked the God of Israel, the chosen dynasty of David, and the chosen city of Jerusalem in a henceforth indissoluble union.

The chosen dynasty and the chosen city

David planned to erect a temple to house the ark, but the tenacious tradition of the ark's portability in a tent shrine forced postponement of the project to his son Solomon's reign. As part of his extensive building operations, Solomon built the Temple on a Jebusite threshing floor, located on a hill north of the royal city, which David had purchased to mark the spot where a plague had been halted. The ground plan of the Temple—a porch with two free-standing pillars before it, a sanctuary, and an inner sanctum—followed Syrian and Phoenician sanctuary models. A bronze "sea" resting on bulls and placed in the Temple court has a Babylonian analogue. Exteriorly, the Jerusalem Temple resembled Canaanite and other Middle Eastern religious structures, but there were differences; *e.g.*, no god image stood in the inner sanctum, but rather only the ancient ark and the new large cherubim (hybrid creatures with animal bodies, human or animal faces, and wings) whose wings covered it, symbolizing the presence of YHWH who was enthroned upon celestial cherubim.

Alongside a brief, ancient inaugural poem in I Kings, chapter 8, verses 12–13, an extensive (and, in its present form, later) prayer expresses the distinctively biblical view of the temple as a vehicle of God's providing for his people's needs. Since, strangely, no reference to sacrifice is made, not a trace appears of the standard pagan conception of the temple as a vehicle of man's providing for the gods.

That literature flourished under the aegis of the court is to be gathered from the quality of the preserved narrative of the reign of David, which gives every indication of having come from the hand of a contemporary eyewitness. The royally sponsored Temple must have had a library and a school attached to it (in accord with the universally attested practice of the ancient Middle East), among whose products were not only royal psalms but also such liturgical pieces intended for the common man as eventually found their way into the book of Psalms.

The latest historical allusions in the Torah literature (the Pentateuch) are to the period of the united monarchy; *e.g.*, the defeat and subjugation of the peoples of Amalek, Moab, and Edom by Saul and David, in Numbers, chapter 24, verses 17–20. On the other hand, the polity reflected in the laws is tribal and decentralized, with no bureaucracy. Its economy is agricultural and pastoral, class distinctions apart from slave and free are lacking, and commerce and urban life are rudimentary. A premonarchic background is evident, with only rare explicit reflections of the later monarchy; *e.g.*, in Deuteronomy, chapter 17, verses 14–20. The groundwork of the Torah literature may thus be supposed to have crystallized under the united monarchy.

The emergence of Torah and wisdom literature

It was in this period that the traditional wisdom cultivated among the learned in neighbouring cultures came to be prized in Israel. Solomon is represented as the author of an extensive literature comparable to that of other Eastern sages. His wisdom is expressly attributed to YHWH in the account of his night oracle at Gibeon (in which he asked not for power or riches but for wisdom), thus marking the adaptation to biblical thought of this common Middle Eastern genre. As set forth in Proverbs, chapter 2, verse 5, "It is YHWH who grants wisdom; knowledge and understanding are by his command." Patronage of wisdom literature is ascribed to the later Judahite king, Hezekiah, and the connection of wisdom with kings is common in extrabiblical cultures as well.

Domination of all of Palestine entailed the absorption of "the rest of the Amorites"—the pre-Israelite population that lived chiefly in the valleys and on the coast. Their impact on Israelite religion is unknown, though some scholars contend that a "royally sponsored syncretism" arose with the aim of fusing the two populations. That popular religion did not meet the standards of the biblical writers and that it incorporated pagan elements—and that such elements may have increased as a result of intercourse with the newly absorbed "Amorites"—is likely and required no royal sponsorship. On the other hand, the court itself welcomed foreigners—Philistines, Cretans, Hittites, and Ishmaelites are named, among others—and made use of their service. Their effect on the court religion may be surmised from what is recorded concerning Solomon's many diplomatic marriages: foreign princesses whom Solomon married brought along with them the apparatus of their native cults, and the King had shrines to their gods built and maintained on the Mount of Olives. Such private cults, while indeed royally sponsored, did not make the religion of the people syncretistic.

Such compromise with the pagan world, entailed by the widening horizons of the monarchy, violated the sanctity of the holy land of YHWH and turned the king into an idolator in the eyes of zealots. Religious opposition, combined with grievances against the organization of forced labour for state projects, led to the secession of the northern tribes (headed by the Joseph tribes) after Solomon's death.

THE PERIOD OF THE DIVIDED KINGDOM

Jeroboam I (10th century BCE), the first king of the north (now called Israel, in contradistinction to Judah, the southern Davidic kingdom), appreciated the inextricable link of Jerusalem and its sanctuary with the Davidic claim to divine election to kingship over all Israel (the whole people, north and south). He therefore founded rival sanctuaries at Dan and at Bethel—ancient cult sites—and manned them with non-Levite priests whose symbol of YHWH's presence was a golden calf—a pedestal of divine images in ancient iconography and the equivalent of the cherubim of Jerusalem's Temple. He also moved the autumn ingathering festival one month ahead so as to foreclose celebration of this most popular of all festivals in common with Judah.

Religion in the northern kingdom (Israel)

For the evaluation of Jeroboam's innovations and the subsequent official religion of the north down to the mid-8th century, one must rely almost exclusively on the Book of Kings (later divided into two books). This work has severe limitations as a source for religious history. The material of this book, in good part contemporary, is subjugated to a dogmatic historiography that regards the whole enterprise of the north as one long apostasy ending in a deserved disaster. The culmination of Kings' history with the exile of Judah shows its provenience to have been Judahite. Yet the evaluation of Judah's official religion is subject to an equally dogmatic standard, namely, the royal adherence to the Deuteronomic rule of a single cult site. The author considered the Solomonic Temple to be the cult site chosen by God, according to Deuteronomy, chapter 12, the existence of which rendered all other sites illegitimate. Every king of Judah is judged according to whether or not he did away with all extra-Jerusalemite places of worship. (The date of this criterion may be inferred from the indifference toward it of all persons [*e.g.*, the 9th-century-BCE prophets Elijah and Elisha and the Jerusalemite priest Jehoiada] prior to the late-8th-century-BCE Judahite king Hezekiah.) Another serious limitation is the restriction of Kings' purview: excepting the Elijah–Elisha stories, it notices only the royally sponsored cult; notices of the popular religion are very few. From the mid-8th century the writings of the classical prophets, starting with Amos, set in. These take in the people as a whole, in contrast to Kings; on the other hand their interest in theodicy (justification of God) and their polemical tendency to exaggerate and generalize what they deem evil must be taken into consideration before approving their statements as sober history.

For a half-century after the north's secession (*c.* 922 BCE), the religious situation in Jerusalem was unchanged. The distaff side of the royal household perpetuated, and even augmented, the pagan cults. King Asa (reigned *c.* 908–867 BCE) is credited with a general purge, including the destruction of an image made for the goddess Asherah by the queen mother, granddaughter of an Aramaean princess. He also purged the *qedeshim* ("consecrated men"—conventionally rendered as "sodomites," or "male sacred prostitutes").

Religion in the southern kingdom (Judah)

Foreign cults entered the north with the marriage of the 9th-century-BCE king Ahab to the Tyrian princess Jezebel. Jezebel brought with her a large entourage of sacred personnel to staff the temple of Baal and Asherah that Ahab built for her in Samaria, the capital of the northern kingdom of Israel. In all else, Ahab's orthodoxy was irreproachable, though others of his court may have joined the worship of the foreign princess. That fierce opposition to the non-YWHW cults sprang up must be supposed in order to account for Jezebel's persecution of the prophets of YHWH, conduct untypical of a polytheist except in self-defense. Elijah's assertion that the whole country apostatized is a hyperbole based on the view that whoever did not actively fight Jezebel was implicated in her polluted cult. Such must have been the view of the prophets, whose fallen were the first martyrs to die for the glory of God. The quality of their opposition may be gauged by Elijah's summary execution of the foreign Baal cultists after they failed the test at Mt. Carmel, where they vied against him in a contest over whose god was truly God. A three-year drought (attested also in Phoenician sources), declared by Elijah to be punishment for the sin, must have done much to kindle the prophets' zeal.

To judge from the Elisha stories, the Baal worship in the capital city, Samaria, was not felt in the countryside.

There the religious tone was set by the popular prophets and the prophetic companies ("the sons of the prophets") who attached themselves to them. In popular consciousness these men were wonder-workers—healing the sick and reviving the dead, foretelling the future, and helping to find lost objects. To the biblical narrator they witness the working of God in Israel. Elijah's rage at the Israelite king Ahaziah's recourse to the pagan god Baalzebub, Elisha's cure of the Syrian military leader Naaman's leprosy, and anonymous prophets' directives and predictions in matters of peace and war all serve to glorify God. Indeed, the equation of Israel's prosperity with God's interest generated the issue of "true" and "false" prophecy that made its first appearance at this time. That prophecy of success could turn out to be a snare is exemplified in a story of conflict between Micaiah, the lone 9th-century-BCE prophet of doom, and 400 unanimous prophets of victory who lured Ahab to his death. The poignancy of the issue is highlighted by Micaiah's acknowledgment that the 400 were also prophets of YHWH—but inspired by him deliberately with a "lying spirit."

The issue of "true" and "false" prophecy

THE PERIOD OF CLASSICAL PROPHECY AND CULT REFORM

The emergence of the literary prophets. By the mid-8th century a hundred years of chronic warfare between Israel and Aram had finally ended—the Aramaeans having suffered heavy blows from the Assyrians. King Jeroboam II (8th century BCE) was able to undertake to restore the imperial sway of the north over its neighbour, and a prophecy of Jonah that he would extend Israel's borders from the Dead Sea to the entrance to Hamath (Syria) was borne out. The well-to-do expressed their relief in lavish attentions to the institutions of worship and their private mansions. But the strain of the prolonged warfare showed in the polarization of society between the wealthy few who had profited from the war and the masses whom it had ravaged and impoverished. Dismay at the dissolution of Israelite society animated a new breed of prophets who now appeared—the literary or classical prophets, first of whom was Amos, an 8th-century-BCE Judahite who went north to Bethel.

That apostasy would set God against the community was an old conception of early prophecy; that violation of the sociomoral injunctions of the Covenant would have the same result was first proclaimed by Amos. Amos almost ignored idolatry, denouncing instead the corruption and callousness of the oligarchy and rulers. The religious exercises of such villains he proclaimed were loathsome to God; on their account Israel would be oppressed from the entrance to Hamath to the Dead Sea and exiled from its land.

Prophetic denouncements of social injustices

The westward push of the Neo-Assyrian Empire in the mid-8th century BCE soon brought Aram and Israel to their knees. In 733–732 Assyria took Gilead and Galilee from Israel and captured Aramaean Damascus; in 721 Samaria, the Israelite capital, fell. The northern kingdom sought to survive through alliances with Assyria and Egypt; its kings came and went in rapid succession. The troubled society's malaise was interpreted by Hosea, a prophet of the northern kingdom (Israel), as a forgetting of God. As a result, in his view, all authority had evaporated: the king was scoffed at, priests became hypocrites, and pleasure seeking became the order of the day. The monarchy was godless; it put its trust in arms, fortifications, and alliances with the great powers. Salvation, however, lay in none of these but in repentance and reliance upon God.

Prophecy in the southern kingdom. Judah was subjected to such intense pressure to join an Israelite–Aramaean coalition against Assyria that its 8th-century-BCE king Ahaz chose to submit himself to Assyria in return for relief. Ahaz introduced a new Aramaean-style altar in the Jerusalem Temple and adopted other foreign customs that are counted against him in the book of Kings. It was at this time that Isaiah prophesied in Jerusalem. At first (under Uzziah, Ahaz' prosperous grandfather), his message focussed on the corruption of Judah's society and religion, stressing the new prophetic themes of indifference to God (which went hand in hand with a thriving cult) and the fateful importance of social morality. Under Ahaz the political crisis evoked Isaiah's appeals for trust in God, with the warning that the "hired razor from across the Euphrates" would shave Judah clean as well. Isaiah interpreted the inexorable advance of Assyria as God's chastisement; Assyria was "the rod of God's wrath." But, since Assyria ignored its mere instrumentality and exceeded in an insolent manner its proper function, God, when he finished his purgative work, would break Assyria on Judah's mountains. Then the nations of the world, who had been subjugated by Assyria, would recognize the God of Israel as the lord of history. A renewed Israel would prosper under the reign of an ideal Davidic king, all men would flock to Zion (the hill symbolizing Jerusalem) to learn the ways of YHWH and submit to his adjudication, and universal peace would prevail (see also DOCTRINES AND DOGMAS: *Eschatology*)

The prophecy of Micah (8th century BCE), also a Judahite, was contemporary with that of Isaiah and touched on similar themes (*e.g.*, the vision of universal peace is found in both their books). Unlike Isaiah, however, who believed in the inviolability of Jerusalem, Micah shocked his audience with the announcement that the wickedness of its rulers would cause Zion to become a plowed field, Jerusalem a heap of ruins, and the Temple mount a wooded height. Moreover, from the precedence of social morality over the cult, Micah drew the extreme conclusion that the cult had no ultimate value and that God's requirement of men can be summed up as "to do justice, and to love kindness, and to walk humbly with your God."

Reforms in the southern kingdom. According to Jeremiah (about 100 years later), Micah's prophetic threat to Jerusalem had caused King Hezekiah (reigned *c.* 715–*c.* 686 BCE) to placate God—possibly an allusion to the cult reform instituted by the King in order to cleanse Judah from various pagan practices. A heightened concern over assimilatory trends resulted in his also outlawing certain practices considered legitimate up to his time. Thus, in addition to removing the bronze serpent that had been ascribed to Moses (and that had become a fetish), the reform did away with the local altars and stone pillars, the venerable (patriarchal) antiquity of which did not save them from the taint of imitation of Canaanite practice. Hezekiah's reform, part of a restorational policy that had political, as well as religious, implications, appears as the most significant effect of the fall of the northern kingdom on official religion. The outlook of the reformers is suggested by the catalog in II Kings, chapter 17, of religious offenses that had caused the fall, which the objects of Hezekiah's purge closely resemble. Hezekiah's reform is the first historical evidence for Deuteronomy's doctrine of cult centralization. Similarities between Deuteronomy and the Book of Hosea lend colour to the supposition that the reform movement in Judah, which culminated a century later under King Josiah, was sparked by attitudes inherited from the north.

The reform of King Hezekiah

Hezekiah was the leading figure in a western coalition of states that coordinated a rebellion against the Assyrian king Sennacherib with the Babylonian rebel Merodach-Baladan, shortly after the Assyrian's accession in 705 BCE. When Sennacherib appeared in the west in 701, the rebellion collapsed; Egypt sent a force to aid the rebels, but it was defeated. Hezekiah saw his kingdom overwhelmed and offered tribute to Sennacherib; the Assyrian, however, pressed for the surrender of Jerusalem. In despair, Hezekiah turned to the prophet Isaiah for an oracle. Though the prophet condemned the King's reliance upon Egyptian help, he stood firm in his faith that Jerusalem's destiny precluded its fall into heathen hands. The King held fast, and Sennacherib, for reasons still obscure, suddenly retired from Judah and returned home. This unlooked-for deliverance of the city may have been regarded as a vindication of the prophet's faith and was doubtless an inspiration to the rebels against Babylonia a century later. For the present, while Jerusalem was intact, the country had been devastated and its kingdom turned into a vassal state of Assyria.

During Manasseh's long and peaceful reign in the 7th century BCE, Judah was a submissive ally of Assyria. Manasseh's forces served in the building and military

Foreign influences and the Deuteronomic reform

operations of the Assyrian kings Esarhaddon and Ashurbanipal. Judah benefitted from the upsurge of commerce that resulted from the political unification of the whole Near East. The prophet Zephaniah attests to heavy foreign influence on the mores of Jerusalem—merchants who adopted foreign dress, cynics who lost faith in the efficacy of YHWH to do anything, people who worshipped the pagan host of heaven on their roofs. Manasseh's court was the centre of such influences. The royal sanctuary became the home of a congeries of foreign gods—the sun, astral deities, and Asherah (the female fertility deity) all had their cults there alongside YHWH. The countryside also was provided with pagan altars and priests, alongside the local YHWH altars that were revived. Presumably, at least some of the blood that Manasseh is said to have spilled freely in Jerusalem must have belonged to YHWH's devotees. No prophecy is dated to his long reign.

With Ashurbanipal's death in 627, Assyria's power faded quickly; the young Judahite king Josiah (reigned *c.* 640–609 BCE) had already set in motion a vigorous movement of independence and restoration, a cardinal aspect of which was religious. First came the purge of foreign cults in Jerusalem, under the aegis of the high priest Hilkiah; then the countryside was cleansed. In the course of renovating the Temple, a scroll of Moses' Torah (by scholarly consensus an edition of Deuteronomy) was found. Anxious to abide by its injunctions, Josiah had the local YHWH altars polluted to render them unusable and collected their priests in Jerusalem. The celebration of the Passover that year was concentrated in the Temple, as it had not been "since the days of the judges who judged Israel," according to II Kings 23:22, or since the days of Samuel, according to II Chron. 35:18; both references reflect the unhistorical theory of the Deuteronomic (Josianic) reformers that the Shiloh sanctuary was the precursor of the Jerusalem Temple as the sole legitimate site of worship in Israel (as demanded by Deuteronomy, chapter 12). To seal the reform, the King convoked a representative assembly and had them enter into a covenant with God over the new-found Torah. For the first time, the power of the state was enlisted on behalf of the ancient covenant and in obedience to a covenant document. It was a major step toward the fixation of a sacred canon.

Josiah envisaged the restoration of Davidic authority over the entire domain of ancient Israel, and the retreat of Assyria facilitated his program—until he became fatally embroiled in the struggle of the powers over the dying empire. His death in 609 was doubtless a setback for his religious policy as well as his political aspirations. To be sure, the royally sponsored syncretism of Manasseh's time was not revived, but there is evidence of recrudescence of unofficial local altars. Whether references in Jeremiah and Ezekiel to child sacrifice to YHWH reflect post-Josianic practices is uncertain. There is stronger indication of private recourse to pagan cults in the worsening political situation.

That Assyria's fall should have been followed by the yoke of a harsh new heathen power dismayed the devotees of YHWH who had not been prepared for it by prophecy. Their mood finds expression in the oracles of the prophet Habakkuk in the last years of the 7th century BCE. Confessing perplexity at God's toleration of the success of the wicked in subjugating the righteous, the prophet affirms his faith in the coming salvation of YHWH, tarry though it might. And in the meantime, "the righteous must live in his faith."

Prophecies of Jeremiah and Ezekiel

But the situation in fact grew worse as Judah was caught in the Babylonian–Egyptian rivalry. Some attributed the deterioration to the burden of Manasseh's sin that still rested on the people. For the prophet Jeremiah (active *c.* 626–*c.* 580 BCE), the Josianic era was only an interlude in Israel's career of guilt that went back to its origins. His pre-reform prophecies denounced Israel as a faithless wife and warned of imminent retribution at the hands of a nameless northerner. After Nebuchadrezzar's decisive defeat of Egypt at Carchemish (605 BCE), Jeremiah identified the scourge as Babylon. King Jehoiakim's attempt to be free of Babylonia ended with the exile of his successor, Jehoiachin, along with Judah's elite (597); yet the court of the new king, Zedekiah, persisted in plotting new revolts, relying—against all experience—on Egyptian support. Jeremiah now proclaimed a scandalous doctrine of the duty of all nations, Judah included, to submit to the divinely appointed world ruler, the Babylonian monarch Nebuchadrezzar. In submission lay the only hope of avoiding destruction; a term of 70 years had been set to humiliate all men beneath Babylon. Imprisoned for demoralizing the populace, Jeremiah persisted in what was viewed as his traitorous message; the leaders, on their part, persisted in their policy, confident of Egypt and the saving power of Jerusalem's Temple, to the bitter end.

Jeremiah also had a message of comfort for his hearers. He foresaw the restoration of the entire people—north and south—in the land, under a new David. And since events had shown that man was incapable of achieving a lasting reconciliation with God on his own, he envisioned the penitent of the future being met halfway by God, who would remake their nature so that to do his will would come naturally to them. God's new covenant with Israel would be written on their hearts, so that they should no longer need to teach each other obedience, for young and old would know YHWH.

Among the exiles in Babylonia, the prophet Ezekiel, Jeremiah's contemporary, was haunted by the burden of Israel's sin. He saw the defiled Temple of Manasseh's time as present before his eyes, and described God as abandoning it and Jerusalem to their fates. Though Jeremiah offered hope through submission, Ezekiel prophesied an inexorable, total destruction as the condition of reconciliation with God. The majesty of God was too grossly offended for any lesser satisfaction. The glory of God demanded Israel's ruin, but the same cause required its restoration. For Israel's fall disgraced YHWH among the nations; to save his reputation he must therefore restore Israel to its land. The dried bones of Israel must revive, that they and all the nations should know that he was YHWH (Ezek. 37). Ezekiel, too, foresaw the remaking of human nature, but as a necessity of God's glorification; the concatenation of Israel's sin, exile, and consequent defamation of God's name must never be repeated. In 587/586 BCE the doom prophecies of Jeremiah and Ezekiel came true. Rebellious Jerusalem was reduced by Nebuchadrezzar, the Temple was burnt, and much of Judah's population dispersed or deported to Babylonia.

THE EXILIC PERIOD

Postexilic religious forms

The survival of the religious community of exiles in Babylonia demonstrates how rooted and widespread the religion of YHWH was. Abandonment of the national religion as an outcome of the disaster is recorded of a minority only. There were some cries of despair, but the persistence of prophecy among the exiles shows that their religious vitality had not flagged. The Babylonian Jewish community, in which the cream of Judah lived, had no sanctuary or altar (in contrast to the Jewish garrison of Elephantine in Egypt); what developed in their place can be surmised from new postexilic religious forms: fixed prayer; public fasts and confessions; and assembly for the study of the Torah, which may have developed from visits to the prophets for oracular edification. The absence of a local or territorial focus must also have spurred the formation of a literary–ideational centre of communal life—the sacred canon of Covenant documents that came to be the core of the present Pentateuch. Observance of the Sabbath—a peculiarly public feature of communal life—achieved a significance among the exiles virtually equivalent to all the rest of the Covenant rules together. Notwithstanding its political impotence, the spirit of the exiles was so high that foreigners were attracted to their ranks, hopeful of sharing their future glory.

Assurance of that future glory was given not only in the consolations promised by Jeremiah and Ezekiel (the fulfillment of whose prophecies of doom lent credit to their consolations); the great comforter of the exile was the writer or writers of what is known as Deutero-Isaiah (Isa. 40–66), who perceived in the rise and progress (from *c.* 550) of the Persian king Cyrus II the Great the instrument of God's salvation. Going beyond the national hopes of

Ezekiel, animated by the universal spirit of the pre-exilic Isaiah, Deutero-Isaiah saw in the miraculous restoration of Israel a means of converting the whole world to faith in Israel's God. Israel would thus serve as "a light for the nations, that YHWH's salvation may reach to the end of the earth." In his conception of the vicarious suffering of God's servant—through which atonement is made for the ignorant heathen—Deutero-Isaiah found a handle by which to grasp the enigma of faithful Israel's lowly state among the Gentiles. The idea was destined to play a decisive role in the self-understanding of the Jewish martyrs of the Syrian king Antiochus IV Epiphanes' persecution in the 2nd century BCE (in, for example, Daniel) and later again in the Christian appreciation of the death of Jesus.

THE PERIOD OF THE RESTORATION

After conquering Babylon, Cyrus so far justified the hopes put in him that he allowed those Jews who wished to do so to return and rebuild their Temple. Though, in time, some 40,000 made their way back, they were soon disillusioned by the failure of the glories of the restoration to materialize and by the controversy with the Samaritans, and left off building the Temple. (The Samaritans were a judaized mixture of native north Israelites and Gentile deportees settled by the Assyrians in the erstwhile northern kingdom.) A new religious inspiration came under the governorship of Zerubbabel, a member of the Davidic line, who became the centre of messianic expectations during the anarchy attendant upon the accession to the Persian throne of Darius I (522). The prophets Haggai and Zechariah perceived the disturbances as heralds of an imminent overthrow of the heathen Persian Empire and a worldwide manifestation of God and glorification of Zerubbabel. Against that day they urged the people quickly to complete the building of the Temple. The labour was resumed and completed in 516; but the prophecies remained unfulfilled. Zerubbabel disappears from the biblical narrative, and the spirit of the community flagged again.

Role of repentance and the Torah

The one religious constant in the vicissitudes of the restored community was the mood of repentance and the desire to win back God's favour by adherence to his Covenant rules. The anxiety that underlay this mood produced a hostility to strangers, which encouraged a lasting conflict with the Samaritans, who asked permission to take part in rebuilding the Temple of the God they too worshipped. The Jews, however, rejected them on ill-specified grounds—apparently ethnoreligious; *i.e.*, they felt the Samaritans to be alien to their historical community of faith, especially to its messianic hopes. Nonetheless, intermarriage occurred and precipitated a new crisis when, in 458, the priest Ezra arrived from Babylon, intent on enforcing the regimen of the Torah. By construing ancient and obsolete laws excluding Canaanites and others so as to make them apply to their own times and neighbours, the leaders of the Jews brought about the divorce and expulsion of several dozen non-Jewish wives and their children. Tension between the xenophobic (fear of strangers) and xenophilic (love of strangers) in postexilic Judaism was finally resolved some two centuries later with the development of a formality of religious conversion, whereby Gentiles who so wished could be taken into the Jewish community by a single, simple procedure.

The decisive constitutional event of the new community was the covenant subscribed to by its leaders in 444, making the Torah the law of the land: a charter granted by the Persian king Artaxerxes I to Ezra—scholar and priest of the Babylonian Exile—empowered him to enforce the Torah as the imperial law for the Jews of the province Avar-nahra (Beyond the River), in which the district of Judah (now reduced to a small area) was located. The charter required the publication of the Torah and the publication, in turn, entailed its final editing—now plausibly ascribed to Ezra and his circle. Survival in the Torah of patent inconsistencies and disaccords with the postexilic situation indicate that its materials were by then sacrosanct, to be compiled but no longer created. But these survivals made necessary the immediate invention of a harmonizing and creative method of text interpretation to adjust the Torah to the needs of the times. The Levites were trained in the art of interpreting the text to the people; the first product of the creative exegesis later known as Midrash is to be found in the covenant document of Nehemiah, chapter 9—every item of which shows development, not reproduction, of a ruling of the Torah. Thus, with the publication of the Torah as the law of the Jews the basis of the vast edifice of the Oral Law characteristic of Judaism was laid.

Origins of the Oral Law

Concern over observance of the Torah was fed by the gap between messianic expectations and the gray reality of the restoration. The gap signified God's continued displeasure, and the only way to regain his favour was to do his will. Thus it is that Malachi, the last of the prophets, concludes with an admonition to be mindful of the Torah of Moses. God's displeasure, however, had always been signalized by a break in communication with him. As time passed and messianic hopes remained unfulfilled, the sense of a permanent suspension of normal relations with God took hold, and prophecy died out. God, it was believed, would some day be reconciled with his people, and a glorious revival of prophecy would then occur. For the present, however, religious vitality expressed itself in dedication to the development of institutions that would make the Torah effective in life. The course of this development is hidden from view by the dearth of sources from the Persian period. But the community that emerged into the light of history in Hellenistic times is one made over radically by this momentous, quiet process. (Mo.Gr.)

Hellenistic Judaism (4th century BCE–2nd century CE)

THE GREEK PERIOD (332–63 BCE)

Hellenism and Judaism. Actual contact between Greeks and Semites goes back to Minoan and Mycenaean times and is reflected in certain terms in Homer and in other early Greek authors. It is not until the end of the 4th century, however, that Jews are first mentioned by Greek writers, who praise the Jews as brave, self-disciplined, and philosophical.

After being conquered by Alexander the Great (332 BCE), Palestine became part of the Hellenistic kingdom of Ptolemaic Egypt, the policy of which was to permit the Jews considerable cultural and religious freedom.

Hellenizing under Seleucid rule

When in 198 BCE Palestine was conquered by King Antiochus III (247–187 BCE), of the Syrian Seleucid dynasty, the Jews were treated even more liberally, being granted a charter to govern themselves by their own constitution, namely, the Torah. Greek influence, however, was already becoming manifest. Some of the 29 Greek cities of Palestine attained a high level of culture. The mid-3rd century-BCE Zenon papyri—containing the correspondence of a business manager of a high Ptolemaic official—present the picture of a wealthy Jew, Tobiah, who through commercial contact with the Ptolemies acquired a veneer of Hellenism, to judge at least from the pagan and religious expressions in his Greek letters. His son and especially his grandsons became ardent Hellenists. It has been argued that the Hellenic influence was so strong among the Jews of Judaea by the beginning of the 2nd century that if the process had continued without the forcible intervention of the Seleucids in Jewish affairs (see below) Judaean Judaism would have become even more syncretistic than that of Philo, the Hellenistic Jewish philosopher of Alexandria (*c.* 15 BCE–*c.* 40 CE). The apocryphal writer Jesus ben Sirach so bitterly denounced the Hellenizers in Jerusalem (*c.* 180 BCE) that he was forced by the authorities to temper his words.

In the early part of the 2nd century BCE, Hellenizing Jews came into control of the high priesthood itself. Jason as high priest (175–172 BCE) established Jerusalem as a Greek city, Antioch-at-Jerusalem, with Greek educational institutions. His ouster by an even more extreme Hellenizing faction, which established Menelaus (died 162 BCE) as high priest, occasioned a civil war, with the wealthy aristocrats supporting Menelaus and the masses Jason. The Syrian king Antiochus IV Epiphanes, who had initially bestowed exemptions and privileges upon the Jews, intervened upon the request of Menelaus' party. Antiochus' promulgation of decrees against the practice of Judaism and the offensive

The Maccabean revolt

and cruel measures to enforce them led to the revolt of an old priest, Mattathias, and his five sons—the so-called Maccabees or Hasmoneans. It has been conjectured that one of the Dead Sea Scrolls, the *War of the Sons of Light Against the Sons of Darkness*, mirrors the fierceness of this struggle. In any case, the figure of the martyr, as known in Judaism and Christianity—the person who bears witness to the faith through his suffering and death—dates from this event.

The tactics employed both in the countryside and in Jerusalem by the Hasmoneans in their counterattack against Hellenizing Jews, whose children they forcibly circumcised, indicate the inroads that Hellenism had already made. On the whole, however, the chief strength of the Hellenizers lay among the wealthy urban population, while the Maccabees derived their strength from the peasants and urban masses. Yet, there is evidence that the ruthlessness exhibited by the Hasmoneans toward the Greek cities of Palestine had political rather than cultural origins, and that, in fact, they were fighting for personal power no less than for the Torah. In any case, some of those who fought on the side of the Maccabees were idol-worshipping Jews. The Maccabees soon found a modus vivendi with Hellenism: Jonathan (160–142), according to the Jewish historian Josephus (*c.* 38–*c.* 100 CE), negotiated a treaty of friendship with Sparta; Aristobulus (104–103 BCE) actually called himself Philhellene (a lover of Hellenism); Alexander Jannaeus (103–76) hired Greek mercenaries and inscribed his coins with Greek as well as with Hebrew. The Greek influence reached its height under King Herod I of Judaea (37–4 BCE), who built a Greek theatre, amphitheatre, and hippodrome in or near Jerusalem.

Social, political, and religious divisions. During the Hellenistic period the priests were both the wealthiest class and the strongest political group among the Jews of Jerusalem. The wealthiest of all were the Oniad family, who held the hereditary office of high priest until they were replaced by the Hasmoneans; the Temple that they supervised was, in effect, a bank, where the Temple wealth was kept and where private individuals also deposited their money. Hence, from a social and economic point of view, Josephus is justified in calling the government of Judaea a theocracy (rule by those having religious authority). Opposition to the priests' oppression arose among an urban middle class group known as scribes (*soferim*), who were interpreters and instructors of the Torah on the basis of an oral tradition probably going back to the time of the return from the Babylonian Exile (538 BCE and after). A special group of the scribes known as Ḥasidim (Greek, Hasideans), or "Pietists," became the forerunners of the Pharisees (middle-class liberal Jews who reinterpreted the Torah and the prophetic writings to meet the needs of their times) and joined the Hasmoneans in the struggle against the Hellenists, though on religious rather than on political grounds.

Josephus held that the Pharisees and the other Jewish parties were philosophical schools, and some modern scholars have argued that the groupings were primarily along economic and social lines; but the chief distinctions among them were religious and go back well before the Maccabean revolt. The equation of Pharisaic with "normative" Judaism can no longer be supported, at any rate not before the destruction of the Temple in 70 CE. The fact that in 70 CE, according to the Palestinian Talmud (see below *Rabbinic Judaism*), there were 24 types of "heretics" in Palestine indicates that there was, in fact, much divergence among Jews; and this picture is confirmed by Josephus, who notes numerous instances of religious leaders who claimed to be prophets and who obtained considerable followings.

The Pharisees and the Sadducees

Some other modern scholars have sought to interpret the Pharisees' opposition to the Sadducees—wealthy, conservative Jews who accepted the Torah alone as authoritative—as based on an urban–rural dichotomy; but a very large share of Pharisaic concern was with agricultural matters. To associate the rabbis with urbanization seems a distortion. The chief support for the Pharisees came from the lower classes, whether in the country or in the city.

The chief doctrine of the Pharisees (literally "Separatists") was that the Oral Law had been revealed to Moses at the same time as the Written Law. In their exegesis and interpretation of this oral tradition, particularly under the rabbi Hillel at the end of the 1st century BCE, the Pharisees were liberal, and their regard for the public won them considerable support. That the Maccabean ruler John Hyrcanus I broke with them and that Josephus set their number at merely "more than 6,000" at the time of King Herod indicates that they were less numerous and influential than Josephus would have his readers believe. The Pharisees stressed the importance of performing all the commandments, including those that appeared to be of only minor significance; those who were particularly strict in their observance of the Levitical rules were known as *ḥaverim* ("companions"). They believed in the providential guidance of the universe, in angels, in reward and punishment in the world to come, and in resurrection of the dead, in all of which beliefs they were opposed by the Sadducees. In finding a modus vivendi with Hellenism, at least in form and in terminology, however, the Pharisees did not differ greatly from the Sadducees. Indeed, the supreme council of the Great Synagogue (or Great Assembly) of the Pharisees was modelled in its organization on Hellenistic religious and social associations. Because they did not take an active role in fostering the rebellion against Rome in 66–70 CE, they were able, through their leader Johanan ben Zakkai, to obtain Roman permission to establish an academy at Jabneh (Jamnia), where, in effect, they replaced the cult of the Temple with study and prayer.

The Sadducees and their subsidiary group, the Boethusians (Boethosaeans), who were identified with the great landowners and priestly families, were more deeply influenced by Hellenization. The rise of the Pharisees may thus be seen, in a sense, as a reaction against the more profound Hellenization favoured by the Sadducees, who were allied with the philhellenic Hasmoneans. From the time of John Hyrcanus (135–104 BCE) the Sadducees generally held a higher position in comparison with the Pharisees and were in favour with the Jewish rulers. Religiously more conservative than the Pharisees, they rejected the idea of a revealed oral interpretation of the Torah, though, to be sure, they had their own tradition, the *sefer gezerot* ("book of decrees" or "decisions"). They similarly rejected the inspiration of the prophetic books of the Bible, as well as the Pharisaic beliefs in angels, rewards, and punishments in the world to come, providential governance of human events, and resurrection of the dead. For them Judaism centred on the Temple; but about 10 years before the destruction of the Temple in 70 CE, the Sadducees in effect disappeared from Jewish life when the Pharisees excluded them from entering the Temple.

The "people of the land" and converts

Not constituting any particular party were the unlearned rural masses known as *ʿamme ha-aretz* ("people of the land"), who were to be found among both the Pharisees and Sadducees and even among the Samaritans, descendants of the northern Israelites who had their own Torah and their own sanctuary. The *ʿamme ha-aretz* did not give the prescribed tithes, did not observe the laws of purity, and were neglectful of the laws of prayer; and so great was the antagonism between them and the learned Pharisees that to their daughters was applied the biblical verse, "Cursed be he who lies with any kind of beast." The antipathy was reciprocated, for in the same passage in the Babylonian Talmud (*Pesaḥim*) are added the words, "Greater is the hatred wherewith the *ʿamme ha-aretz* hate the scholar than the hatred wherewith the heathens hate Israel." That there was, however, social mobility is clear from the Talmudic dictum, "Heed the sons of the *ʿam ha-aretz*, for they will be the living source of the Torah." That there is little evidence that the early Christian church was particularly successful in converting *ʿamme ha-aretz* suggests that their position was not unbearable.

Proselytes (converts) to Judaism, though not constituting a class, became increasingly numerous both in Palestine and especially in the Diaspora (the Jews living beyond Palestine). Scholarly estimates of the Jewish population of this era range from 700,000 to 5,000,000 in Palestine and from 2,000,000 to 5,000,000 in the Diaspora, with the prevailing opinion being that about one-tenth of the population of the Mediterranean world at the beginning of the Christian Era was Jewish. Such numbers represent a considerable increase from previous eras and must have included large numbers of proselytes. Already in 139 BCE the Jews of Rome were charged by the praetor (civil administrator) with attempting to contaminate Roman morals with their religion, presumably an allusion to proselytism. The first large-scale conversions were by John Hyrcanus and Aristobulus, who, in 130 and 103 BCE, respectively, forced the people of Idumaea in southern Palestine and of Ituraea in northern Palestine to become Jews. The eagerness of the Pharisees to win converts is seen in a statement in Matthew that the Pharisees would "traverse sea and land to make a single proselyte." To be sure, some of the proselytes, according to Josephus, did return to their pagan ways, but the majority apparently remained true to their new religion. In addition, there were many "sympathizers" with Judaism who observed one or more Jewish practices without being fully converted.

Outside the pale of Judaism in most, though not all, respects were the Samaritans, who, like the Sadducees, refused to recognize the validity of the Oral Law; and, in fact, the break between the Sadducees and the Samaritans did not occur until the conquest of Shechem by John Hyrcanus (128 BCE). Like the later so-called Qumrān covenanters (the monastic group with whom are associated the Dead Sea Scrolls), they were opposed to the Jewish priesthood and the cult of the Temple, regarded Moses as a messianic figure, and forbade the revelation of esoteric doctrines to outsiders.

Scholars have recently revised an older conception of a "normative" Pharisaic Judaism dominant in Palestine and a deviant Judaism dominant in the Diaspora. On the one hand, the picture of "normative" Judaism is broader than at first believed, and it is clear that there were many differences of emphasis within the Pharisaic party; and, on the other hand, supposed differences between Alexandrian and Palestinian Judaism were not as great as had been formerly thought. In Palestine, no less than in the Diaspora, there were then deviations from Pharisaic standards.

Hellenistic influences on Pharisaic Judaism

Despite the attempts of the Pharisaic leaders to restrain the wave of Greek influence, they themselves showed at least a surface Hellenization. In the first place, as many as 2,500–3,000 words of Greek origin are to be found in the Talmudic corpus, and they supply important terms in the fields of law, government, science, religion, technology, and everyday life, especially in the popular sermons preached by the rabbis. When preaching, the Talmudic rabbis often gave the Greek translation of biblical verses for the benefit of those who understood Greek only. The prevalence of Greek in ossuary (burial) inscriptions and the discovery of Greek papyri in the Dead Sea caves confirm the widespread use of the language, though few Jews, it seems, really mastered Greek. Again, there was a surface Hellenization in the frequent adoption of Greek names, even by the rabbis; and there is evidence (Talmud, *Soṭa*) of a school at the beginning of the 2nd century that had 500 students of "Greek wisdom." Even after 117 CE, when it was prohibited by the rabbis to teach one's son Greek, Rabbi Judah the Prince, the editor of the Mishna (authoritative compilation of the Oral Law) at the end of the 2nd century, remarked, "Why talk Syriac in Palestine? Talk either Hebrew or Greek." Even the synagogues of the period have the form of Hellenistic-Roman basilicas, have frequent inscriptions in Greek, and often have pagan motifs. Many of the anecdotes told about the rabbis have Socratic and Cynic parallels. There is evidence of discussions of rabbis with Athenians, Alexandrians, and Roman philosophers, and even with the emperor Antoninus; but in all of these discussions there is evidence of only one rabbi, Elisha ben Abuyah, who became a Gnostic heretic, accepting certain esoteric religious dualistic views. The rabbis never mention the Greek philosophers Plato or Aristotle or the Hellenistic Jewish philosopher Philo, and they never use any Greek philosophical terms; the only Greek author whom they name is Homer. Again, the parallels between Hellenistic rhetoric and rabbinic hermeneutics are in the realm of terminology rather than of substance, and those between Roman and Talmudic law are inconclusive. Part of the explanation of this may be that, although there were 29 Greek cities in Palestine, none was in Judaea, the real stronghold of the Jews.

Important historical sites of Hellenistic and medieval Judaism.

Religious rites and customs in Palestine: Temple and synagogues. The most important religious institution of the Jews until its destruction in 70 was the Temple in Jerusalem—the Second Temple, erected 538–516 BCE. Though services were interrupted for three years by Antiochus Epiphanes (167–165 BCE) and though the Roman general Pompey desecrated the Temple (63 BCE), Herod lavished great expense in rebuilding it. The high priesthood itself became degraded by the extreme Hellenism of such high priests as Jason and Menelaus; and the institution declined when Herod began the custom of appointing the high priests for political and financial considerations. That not only the multitude of Jews but the priesthood itself suffered from sharp divisions is clear from the bitter class warfare that ultimately erupted in 59 CE between the high priests on the one hand and the ordinary priests and the leaders of the populace of Jerusalem on the other.

Though the Temple remained central in Jewish worship, synagogues may already have emerged during the Babylonian Exile in the 6th century BCE. In any case, in the following century, Ezra stood upon a pulpit of wood and read from the Torah to the people (Nehemiah). According to the interpretation of some scholars, a synagogue existed even within the precincts of the Temple; and certainly by the time of Jesus, to judge from the references to Galilean synagogues in the New Testament, synagogues were common in Palestine. Hence, when the Temple was destroyed in 70, the spiritual vacuum was hardly as great as it had been after the destruction of the First Temple (586 BCE).

The Sanhedrin

The chief legislative, judicial, and educational body of the Palestinian Jews during the period of the Second Temple was the Great Sanhedrin (council court), consisting of 71 members, among whom the Sadducees were an important party. The members shared the government with the king during the early years of the Hasmonean dynasty, but beginning with Herod's reign their authority was restricted to religious matters. In addition, there was another Sanhedrin, set up by the high priest, which served as a court of political council, as well as a kind of grand jury.

Religious and cultural life in the Diaspora. During the Hellenistic-Roman period the chief centres of Jewish population outside Palestine were in Syria, Asia Minor, Babylonia, and Egypt, each of which is estimated to have had at least 1,000,000 Jews. The large Jewish community of Antioch—which, according to Josephus, had been given all the rights of citizenship by the Seleucid founder-king, Seleucus Nicator (died 280 BCE)—attracted a particularly large number of converts to Judaism. It was in Antioch that the apocryphal book of Tobit was probably composed in the 2nd century BCE to encourage wayward Diaspora Jews to return to their Judaism. As for the Jews of Asia Minor, whose large numbers were mentioned by Cicero (1st century BCE), their not joining in the Jewish revolts against the Roman emperors Nero, Trajan, and Hadrian would indicate that they had sunk deep roots into their environment. In Babylonia, in the early part of the 1st century CE, two Jewish brothers, Asinaeus and Anilaeus, were able to establish an independent minor state; their followers were so meticulous in observing the Sabbath that they assumed that it would not be possible to violate the Sabbath even in order to save themselves from a Parthian attack. In the early part of the 1st century CE, according to Josephus, the royal house and many of their entourage in the district of Adiabene in northern Mesopotamia were converted to Judaism; some of the Adiabenian Jews distinguished themselves in the revolt against Rome in 66 (see below).

The Egyptian Diaspora

The largest and most important Jewish settlement in the Diaspora was in Egypt. There is evidence (papyri) of a Jewish military colony at Elephantine (Yeb), Upper Egypt, as early as the 6th century BCE. These papyri reveal the existence of a Jewish temple—which most certainly would be considered heterodox—and some syncretism (mixture) with pagan cults. Alexandria, the most populous and most influential Hellenistic Jewish community in the Diaspora, had its origin when Alexander the Great assigned a quarter of the city to the Jews. Until about the 3rd century BCE the papyri of the Egyptian Jewish community were written in Aramaic; after that, with the exception of the Nash papyrus in Hebrew, all papyri until 400 CE were in Greek. Similarly, of the 116 Jewish inscriptions from Egypt, all but five are written in Greek. The process of Hellenistic acculturation is, thus, obvious.

The most important work of the early Hellenistic period, dating, according to tradition, from the 3rd century BCE, is the Septuagint, a translation of the Pentateuch into Greek. (The translation of the whole Hebrew Bible was completed during the next two centuries.) The fact that, in the *Letter of Aristeas* (see below) and the works of Philo and Josephus, this translation was itself regarded as divinely inspired led to the neglect of the Hebrew original. The translation shows some knowledge of Palestinian exegesis and the tradition of Halakha (the Oral Law); but the rabbis themselves, noting that the translation diverged from the Hebrew text, apparently had ambivalent feelings about it, as is evidenced in their alternate praise and condemnation of it. The fact that such a concept as Torah was translated as *nomos* ("law") and *tzedaqa* as *dikaiosynē* ("justice") opened the way to antilegalism in early Christianity and to Platonic interpretations; and the introduction of such Greek mythological terms as "Titans" and "Sirens" helped to pave the way for the syncretism of Judaism and paganism.

The establishment of a temple at Leontopolis in Egypt (*c.* 145 BCE) by a deposed high priest, Onias IV, indicates that the temple was clearly heterodox; but this temple never really offered a challenge to the one in Jerusalem and was merely the temple of the military colony of Leontopolis. It is significant that the Palestinian rabbis ruled that a sacrifice intended for the temple of Onias might be offered in Jerusalem. That the temple of Onias made little impact upon Egyptian Jewry can be seen from the silence about it on the part of Philo, who often mentions the Temple in Jerusalem. The temple of Onias, however, continued until it was closed by the Roman emperor Vespasian in 73 CE.

The chief religious institutions of the Egyptian Diaspora were synagogues. As early as the 3rd century BCE there were inscriptions mentioning two *proseuchai,* Jewish prayerhouses. In Alexandria there were numerous synagogues throughout the city, of which the largest was so famous that it is said in the Talmud (see below *Rabbinic Judaism*) that he who has not seen it has never seen the glory of Israel.

Egyptian Jewish literature. In Egypt the Jews produced a considerable literature (most of it now lost), intended to inculcate in Greek-speaking Jews a pride in their past and to counteract an inferiority complex that some of them felt about Jewish cultural achievements. In the field of history, Demetrius, near the end of the 3rd century BCE, wrote a work *On the Kings in Judaea*—perhaps intended to refute an anti-Semitic Egyptian priest and author—showing considerable concern for chronology. In the 2nd century BCE a Jew who used the name of Hecataeus wrote *On the Jews.* Another, Eupolemus (*c.* 150 BCE), like Demetrius, wrote *On the Kings in Judaea;* an indication of its apologetic nature may be seen from the fragment asserting that Moses taught the alphabet not only to the Jews but also to the Phoenicians and to the Greeks. Artapanus (*c.* 100 BCE), in his book *On the Jews,* went even further in romanticizing Moses by identifying him with the Greek Musaeus and the Egyptian Hermes-Thoth (god of Egyptian writing and culture) and by asserting that Moses was the real originator of Egyptian civilization and that he even taught the Egyptians the worship of the deity Apis (the sacred bull) and the ibis (sacred bird). In his history, Cleodemus (or Malchus), in an obvious attempt to win for the Jews the regard of the Greeks, asserted that two sons of Abraham had joined Heracles in his expedition in Africa and that the Greek hero had married the daughter of one of them. On the other hand, Jason of Cyrene (*c.* 100 BCE) wrote a history, of which II Maccabees is a summary, glorifying the Temple and violently attacking the Jewish Hellenizers; but his manner of writing history is typically Hellenistic, with emphasis on pathos. III Maccabees (1st century BCE) is a work of propaganda intended to counteract those Jews who sought to win citizenship in Alexandria. The *Letter of Aristeas,* though ascribed to a pagan courtier, Ptolemy II Philadelphus, was probably composed by an Alexandrian

Jew about 100 BCE to defend Judaism and its practices against detractors.

Egyptian Jews also composed poems and plays, now extant only in fragments, to glorify their history. Philo the Elder (*c.* 100 BCE) wrote an epic *On Jerusalem* in Homeric hexameters. Theodotus (*c.* 100 BCE) wrote an epic *On Shechem,* quite clearly apologetic, to judge from the fragment connecting the name of Shechem with Sikimios, the son of the Greek god Hermes. At about the same time, a Jewish poet wrote a didactic poem, ascribing it to the pagan Phocylides, though closely following the Bible in some details; the author disguised his Jewish origin by omitting any attack against idolatry from his moralizing. A collection known as *The Sibylline Oracles,* containing Jewish and Christian prophecies in pagan disguise, includes some material composed by a 2nd-century-BCE Alexandrian Jew who intended to glorify the pious Jews and perhaps to win converts; it is possible that the *Oracles* were known to the Roman poet Virgil when he wrote his fourth *Eclogue.*

A Jewish dramatist of the period, Ezekiel (*c.* 100 BCE), composed tragedies in Greek. Fragments of one of them, *The Exodus,* show how deeply he was influenced by the Greek dramatist Euripides. Whether such plays were actually presented on the stage or not, they edified Jews and showed the pagans that the Jews had as much material for drama as they did.

Alexandrian Jewish philosophical achievements

The greatest achievement of Alexandrian Judaism was in the realm of wisdom literature and philosophy. In a work on the analogical interpretation of the Law of Moses, Aristobulus in the 2nd century BCE anticipated Philo in attempting to harmonize Greek philosophy and the Torah, in using the method of allegory to explain anthropomorphisms in the Bible, and in asserting that the Greek philosophers were indebted to Moses. The Wisdom of Solomon, dating from the 1st century BCE, shows an acquaintance with the Platonic doctrine of the preexistence of the soul and with a method of argument known as *sorites* that was favoured by the Stoics (Greek philosophers). During the same period the author of IV Maccabees showed an intimate knowledge of Greek philosophy, particularly of Stoicism.

By far the greatest figure in Alexandrian Jewish literature is Philo, who has come to be recognized as a major philosopher. His synthesis of Greek philosophy, particularly that of Plato, and of the Torah, and his formulation of the Logos (Word, or Divine Reason) as an intermediary between God and the world, helped lay the groundwork for Neoplatonism (a philosophy dealing with levels of being), Gnosticism (a dualistic religious movement teaching that matter is evil and that spirit is good), and the philosophical framework of the early Church Fathers. Philo was a devotee of Judaism, neither as a mystic cult nor as a collateral branch of Pharisaic Judaism; he was a Diaspora Jew with a profound knowledge of Greek literature who, though almost totally ignorant of Hebrew, tried to find a modus vivendi between Judaism and secular culture.

Mention may be made of the Jewish community of Rome. Numbering perhaps 50,000, it was, to judge from the inscriptions in the Jewish catacombs, predominantly Greek-speaking and almost totally ignorant of Hebrew. References in Roman writers, particularly Tacitus and the satirists, have led scholars to conclude that the community—which was influential, to judge from the pagan jibes—observed the Sabbath and the dietary laws and was active in seeking converts.

The Hellenization of the Diaspora Jews is, however, to be seen not merely in their literature but even more in the papyri and art objects that have recently been studied at great length. As early as 290 BCE, Hecataeus of Abdera, a Greek non-Jew living in Egypt, had remarked that under the Persians and Macedonians the Jews had greatly modified the traditions of their fathers. The fact that—to judge from other papyri—at least three-fourths of the Egyptian Jews had personal names of Greek, rather than Hebrew, origin is significant. That the only schools of which mention is made are Sabbath schools intended for adults and that, on the contrary, Jews were extremely eager to gain admittance for their children to Greek *gymnasia*—where quite obviously they would have to make compromises with their Judaism—indicates their scale of values. Again, there are a number of violations from the norms of Halakha (which precluded the charging of interest for a loan), most notably in the fact that of 11 known extant loan documents only two are without interest. There are often striking similarities between the documents of sale, marriage, and divorce of the Jews and of the Greeks in Egypt, though some of this, as with the documents of the Elephantine Jewish community, may be due to a common origin in the cuneiform law of ancient Mesopotamia. The charms and apotropaic (designed to avert evil) amulets are often syncretistic, and the Jews can hardly have been unaware of the religious significance of symbols that were still very much filled with meaning in pagan cults. The fact that the Jewish community of Alexandria was preoccupied in the 1st century BCE and the 1st century CE with obtaining rights as citizens—which certainly involved compromises with Judaism, including participation in pagan festivals and sacrifices—shows how far they were ready to deviate. Philo mentions Jews who scoffed at the Bible, which they insisted on interpreting literally, and of others who failed to adhere to the biblical laws that they regarded as mere allegory; he writes too of Jews who observed nothing of Judaism except the holiday of Yom Kippur. But despite such deviations, the pagan writers constantly accuse the Diaspora Jews of being "haters of mankind" and of being absurdly superstitious; and Christian writers later similarly attack the Jews for refusing to give up the Torah. At least they were loyal Jews in their contributions of the Temple tax and in pilgrimages to Jerusalem on the three festivals. Actual apostasy and intermarriage were apparently not common, but the virulent anti-Semitism and the pogroms perpetrated by the Egyptian non-Jews must have served as a deterrent.

Modifications of Jewish traditions by Diaspora Jews

Palestinian literature. During this period literature was composed in Palestine in Hebrew, Aramaic, and Greek, with the exact language still a subject of dispute among scholars in many cases and with the works often apparently composed by more than one author over a considerable period of time. Most of the works composed in Hebrew, many of them existing only in Greek—Ecclesiasticus, I Maccabees, Judith, *Testaments of the Twelve Patriarchs,* Baruch, *Psalms of Solomon,* Prayer of Manasseh—and many of the Dead Sea Scrolls are generally conscious imitations of biblical books, often reflecting the dramatic events of the Maccabean struggle and often with an apocalyptic tinge (involving the dramatic intervention of God in history). The literature in Aramaic consists of the following: (1) biblical or Bible-like legends or midrashic (interpretive) additions—*Testament of Job, The Martyrdom of Isaiah, Paralipomena of Jeremiah, Life of Adam and Eve,* the Dead Sea *Genesis Apocryphon,* Tobit, Susanna, Bel and the Dragon; and (2) apocalypses—*Enoch* (perhaps originally written in Hebrew), *Assumption of Moses,* the Syriac Baruch, II (IV) Esdras, and *Apocalypse of Abraham.* In Greek the chief works by Palestinians are histories of the Jewish War against Rome and of the Jewish kings by Justus of Tiberias (both are lost) and the history of the Jewish War, originally in Aramaic, and the *Jewish Antiquities* by Josephus (both written in Rome).

Apocalyptic and wisdom literature

Of the wisdom literature composed in Hebrew, the book of the Wisdom of Jesus the Son of Sirach, or Ecclesiasticus (*c.* 180–175 BCE), modelled on the book of Proverbs, identified Wisdom with the observance of the Torah. The *Testaments of the Twelve Patriarchs,* probably written in the latter half of the 2nd century BCE, patterned on Jacob's blessings to his sons, are now thought to belong to eschatological literature related to the Dead Sea Scrolls. The identification of Wisdom and Torah is stressed in the Mishnaic tract *Pirqe Avot* ("Sayings of the Fathers"), which, though edited 200 CE, contains the aphorisms of rabbis dating back to 300 BCE.

Books such as the *Testament of Job,* the Dead Sea Scroll *Genesis Apocryphon,* the *Book of Jubilees* (now known to have been composed in Hebrew, as seen by its appearance among the Dead Sea Scrolls), and *Biblical Antiquities,* falsely attributed to Philo (originally written in Hebrew, then translated into Greek, but now extant only in Latin), as well as the first half of Josephus' *Jewish Antiquities,* often show affinities with rabbinic Midrashim (interpretive works) in their legendary accretions of biblical details. Sometimes, as in *Jubilees* and in the Pseudo-Philo work, these accretions are intended to answer the questions of heretics, but often, particularly in the case of Josephus, they are apologetic in presenting biblical heroes in a guise that would appeal to a Hellenized audience.

Apocalyptic trends, given considerable impetus by the victory of the Maccabees over the Syrian Greeks, were not—as was formerly thought—restricted to Pharisaic circles. They were (as is clear from the Dead Sea Scrolls) found in other groups as well, and are of particular importance for their influence on both Jewish mysticism and early Christianity. These books, which have a close connection with the biblical Book of Daniel, stress the impossibility of a rational solution to the problem of theodicy—how to reconcile the righteousness of God with observable evil. They also stress the imminence of the day of salvation, which is to be preceded by terrible hardships, and presumably reflected the current historical setting. In the book of *Enoch* there is stress on the terrible punishment inflicted upon sinners in the Last Judgment, the imminent coming of the Messiah and of his kingdom, and the role of angels.

The sole Palestinian Jewish author writing in Greek whose works are preserved is Josephus. His account of the war against the Romans in his *Life* and, to a lesser degree, in the *Jewish War* are largely a defense of his own questionable behaviour as the commander of the Jewish forces in Galilee. But these works and more especially *Against Apion* and the *Jewish Antiquities* are largely defenses of Judaism against anti-Semitic attacks. Josephus' *Jewish War* is often quite deliberately parallel to Thucydides' *History of the Peloponnesian War;* and his *Jewish Antiquities* is quite deliberately parallel to Dionysius of Halicarnassus' *Roman Antiquities,* dating from earlier in the same century.

THE ROMAN PERIOD (63 BCE–135 CE)

New parties and sects. Under Roman rule a number of new groups, largely political, emerged in Palestine. Their common aim was to seek an independent Jewish state. All were zealous for, and strict in their observance of, the Torah.

The Herodians and the Zealots

The Herodians were a political group that after the death of Herod—whom they apparently regarded as the Messiah—sought the reestablishment of the rule of Herod's descendants over an independent Palestine as a prerequisite for Jewish preservation. Unlike the Zealots, however (see below), they did not refuse to pay taxes to the Romans.

The Zealots' party, founded *c.* 6–9 CE, refused to pay tribute to the Romans and advocated overthrowing them on the ground that they should acknowledge God alone as their master. A priestly, eschatologically oriented resistance movement, the Zealots were particularly dedicated to keeping the Temple and its cult pure and used guerrilla tactics toward that end. The Sicarii (Assassins), so-called because of the dagger (*sica*) they carried, arose *c.* 54, according to Josephus, as a group of bandits who kidnapped or murdered those who had found a modus vivendi with the Romans. It was they who made a stand at the fortress of Masada near the Dead Sea, committing suicide rather than be captured by the Romans (73).

The Essenes, New Covenanters, and Dead Sea groups

A number of other parties—various types of Essenes, Damascus Covenanters, and the Qumrān Dead Sea groups—were distinguished by their pursuit of an ascetic monastic life, disdain for material goods and sensual gratification, sharing of material possessions, concern for eschatology, strong apocalyptic views in anticipation of the coming of the Messiah, practice of ablutions to attain greater sexual and ritual purity, prayer, contemplation, and study. The Essenes were like the Therapeutae, a Jewish religious group that had flourished in Egypt two centuries earlier, but the latter actively sought "wisdom" whereas the former were anti-intellectual. Only some of the Essenes were celibate. The Essenes have been termed Gnosticizing Pharisees because of their belief, shared with the Gnostics, that the world of matter was evil; some have seen in them the influence of a quasi-monasticism.

The Damascus sect (New Covenanters) were a group of Pharisees who went beyond the letter of the Pharisaic Halakha. Like the Essenes and the Dead Sea sect, they had a monastic type of organization and opposed the way in which sacrifices were offered in the Temple.

The continuing recent discoveries of scrolls in caves of the Dead Sea area have focussed attention on the groups that lived there. On the basis of paleography, carbon-14 testing, and the coins discovered there, most scholars accept a 1st-century date for them. A theoretical relationship of the communities with John the Baptist and the nascent Christian groups remains in dispute, however. The sectaries have been identified variously as Zealots, an unnamed anti-Roman group, and especially Essenes; but a major difference between the Qumrān groups and the Essenes is that the former were militarily activist (the discovery of hymns and a calendar at Masada—a stronghold of the Sicarii—that had previously been found at Qumrān, may indicate a connection between the groups), while the latter were, for the most part, pacifist. That the groups had secret, presumably apocalyptic, teachings is clear from the fact that among the scrolls are some in cryptographic script and reversed writing; and yet, despite their extreme piety and legalistic conservatism, they apparently were not unaware of Hellenism, to judge from the presence of Greek books at Qumrān.

It has long been debated whether the Gnostic systems of the 1st and 2nd centuries go back to the collapse of the apocalyptic strains in Judaism—which expected a final transforming catastrophic event—when the Temple was destroyed in 70. It is doubtful that there is any direct Jewish source for this Gnosticism, though some characteristic Gnostic doctrines are found in certain groups of particularly apocalyptic 1st-century Jews—the dichotomy of body and soul and a disdain for the material world, a notion of esoteric knowledge, and an intense interest in angels and in problems of creation.

Origin of Christianity: the early Christians and the Jewish community. Though it attracted little attention among pagans and Jews at the beginning, the rise of Christianity was by far the most important "sectarian" development of the Roman period. With the revision, largely due to the discoveries at Qumrān, of the view that Pharisaic Judaism was to be considered normative, primitive Christianity, with its apocalyptic and eschatological interests, has come to be viewed by many scholars as no longer "sectarian" or peripheral to Jewish development but, at least initially, as part of a broad spectrum of attitudes within Judaism. Jesus himself, despite his criticisms of Pharisaic legalism, may now be classified as a Pharisee with strong apocalyptic inclinations; he proclaimed that he had no intention of abrogating the Torah, but of fulfilling it. It is possible to envision a direct line between Jewish currents, both in Palestine and the Diaspora in the Hellenistic Age, and Christianity, particularly in the traditions of martyrdom, proselytism, monasticism, mysticism, liturgy, and religious philosophy, especially the doctrine of the Logos (Word) as an intermediary between God and the world and the synthesis of faith and reason. The Septuagint, in particular, played an important role both theoretically, in the transformation of Greek philosophy into the theology of the Church Fathers, and practically, in converting Jews and Jewish "sympathizers" to Christianity. The connection of nascent Christianity with the Qumrān groups may be seen in their dualism and apocalypticism; but there are differences, notably in the conception of the Incarnation, in the relationship of the Son and the Father, and in Jesus' vicarious suffering for sinners as against the direct suffering of the Qumrān Teacher of Righteousness. Again, the Qumrān group constituted an esoteric movement, militant, with enforced community of goods, concerned with strict observance of the Torah, especially with its calendar, whereas Christianity was pacifist, was open to all, and represented a New Covenant, with stress away from the Torah ritual and with voluntary community of possessions. In general, moreover, Christianity was more positively disposed toward Hellenism than was Pharisaism, particularly under the leadership of Paul, a thoroughly Hellenized Jew.

Primitive Christianity a part of the 1st-century Jewish religious spectrum

When Paul proclaimed his antinomianism (against Torah observance as a means of salvation) many Jewish followers of Jesus became Jewish Christians and continued to observe the Torah. Their two main groupings were the Ebionites—probably to be identified with those called *minim*, or "sectaries," in the Talmud—who accepted Jesus as the Messiah but denied his divinity, and the Nazarenes, who regarded Jesus as both Messiah and God, but regarded the Torah as binding upon Jews alone.

The break between Judaism and Christianity

The percentage of Jews converted to any form of Christianity was extremely small, as can be seen from the frequent criticisms of Jews for their stubbornness by Christian writers. In the Diaspora, despite the strong influence of Hellenism, there were relatively few Jewish converts, though the Christian movement had some success in winning Alexandrian Jews.

There were four major stages in the final break between Christianity and Judaism: (1) the flight of the Jewish Christians from Jerusalem to Pella across the Jordan in 70 and their refusal to continue the struggle against the Romans; (2) the institution by the patriarch Gamaliel II of a prayer in the Eighteen Benedictions against such heretics (*c.* 100), and (3 and 4) the failure of the Christians to join the messianic leaders Lukuas-Andreas and Bar Kokhba in the revolts against Trajan (115–117) and Hadrian (132–135), respectively.

Judaism under Roman rule. When Pompey entered the Temple in 63 BCE as an arbiter both in the civil war between Hyrcanus and Aristobulus and in the struggle of the Pharisees against both Jewish rulers, Judaea in effect became a puppet state of the Romans. During the civil war between Pompey and Julius Caesar, the Idumaean Antipater had ingratiated himself with Caesar by aiding him and was rewarded by being made governor of Judaea; the Jews were rewarded through the promulgation of a number of decrees favourable to them, which were reaffirmed by Augustus and later emperors. His son Herod, king of Judaea, an admirer of Greek culture, supported a cult worshipping the Emperor and built temples to Augustus in non-Jewish cities. Since he was by origin an Idumaean, he was regarded by many Jews as a foreigner. (The Idumaeans, or Edomites, were forcibly converted to Judaism by John Hyrcanus; see above.) On several occasions during and after his reign, Pharisaic delegations sought to convince the Romans to end the quasi-independent Jewish government. After the death of Herod's son and successor Archelaus in 6 CE, his realms were ruled by Roman procurators, the most famous or infamous of whom, Pontius Pilate (26–36), attempted to introduce busts of the Roman emperor into Jerusalem and discovered the intense religious zeal of the Jews in opposing this measure. When Caligula ordered the governor of Syria, Petronius, to install a statue of himself in the Temple, a large number of Jews proclaimed they would suffer death rather than to permit such a desecration. Petronius in response succeeded in getting the Emperor to delay. The procurators of Judaea, being of equestrian (knightly) rank and often of Oriental Greek stock, were more anti-Semitic than the governors of Syria, who were of the higher senatorial order. The last procurators in particular were indifferent to Jewish religious sensibilities; and various patriotic groups, to whom nationalism was an integral part of their religion, succeeded in polarizing the Jewish population and bringing on an extremely bloody war with Rome in 66–70. The climax of the war was the destruction of the Temple in 70, though, according to Josephus, the Roman general (and later emperor) Titus sought to spare it. The war was not ended, however, until 73, when the Sicarii at Masada committed suicide rather than submit to the Romans.

Jewish wars against the Romans

The papyri indicate that the war against Trajan (115–117), involving the Jews of Egypt, Cyrenaica, Cyprus, and Mesopotamia (though only to a minor degree those of Palestine), was a widespread revolt under a Cyrenian king-messiah, Lukuas-Andreas, aimed at freeing Palestine from Roman rule. The same spirit of freedom impelled another messiah, Bar Kokhba, who had the support of the greatest rabbi of the time, Akiba, in his spontaneous uprising (132–135). The result was Hadrian's decrees prohibiting circumcision and public instruction in the Torah, though these were soon revoked by Antoninus Pius. Having suffered such tremendous losses on the field of battle, Judaism turned its dynamism to the continued development of the Talmud (see below *Rabbinic Judaism*). (L.H.F.)

Rabbinic Judaism (2nd–18th centuries)

THE AGE OF THE TANNAIM (135–C. 200)

The role of the rabbis. With the defeat of Bar Kokhba and the ensuing collapse of active Jewish resistance to Roman rule (135–136), politically moderate and quietist rabbinic elements remained the only cohesive group within Jewish society. With Jerusalem off limits to the Jews, rabbinic ideology and practice, which were not dependent on Temple, priesthood, or political independence for their vitality, provided a viable program for autonomous community life and thus filled the vacuum created by the suppression of all other Jewish leadership. The Romans, confident that the will for insurrection had been shattered, soon relaxed the Hadrianic prohibitions of Jewish ordination, public assembly, and regulation of the calendar and permitted rabbis who had fled the country to return and reestablish an academy in the town of Usha in Galilee.

The strength of the rabbinate lay in its ability to represent simultaneously the interests of the Jews and the Romans, whose religious and political needs, respectively, now chanced to coincide. The rabbis were regarded favourably by the Romans, as a politically submissive class, which, with its wide influence over the Jewish masses, could translate the Pax Romana (the peace imposed by Roman rule) into Jewish religious precepts. To the Jews, on the other hand, the rabbinic ideology gave the appearance of continuity to Jewish self-rule and freedom from alien interference. The rabbinic program fashioned by Johanan ben Zakkai's circle (see above *Hellenistic Judaism*) had replaced sacrifice and pilgrimage to the Temple with study of Scripture, prayer, and works of piety, thus eliminating the need for a central sanctuary (in Jerusalem) and making of Judaism a religious association capable of fulfillment anywhere. Judaism was now, for all intents and purposes, a Diaspora religion even on its home soil. Any sense of real break with the past was mitigated by continued adherence to purity laws (dietary and bodily) and by assiduous study of Scripture, including those legal sections that historical developments had now made obsolete. The reward held out for scrupulous study and fulfillment was the promise of messianic deliverance; *i.e.*, divine restoration of all those institutions that had become central in Jewish notions of national independence—the Davidic monarchy, Temple service, the ingathering of Diaspora Jewry—and, above all, the assurance of personal reward to the righteous through resurrection and participation in the national rebirth.

The rabbinic way

Apart from the right to teach Scripture publicly, the most pressing need felt by the surviving rabbis was for the reorganization of a recognized body that would reactivate the functions of the former Sanhedrin and pass on disputed questions of law and dogma. A high court was, accordingly, organized under the leadership of Simeon ben Gamaliel (reigned *c.* 135–*c.* 175), the son of the previous patriarch (the Roman term for the head of the Palestinian Jewish community) of the house of Hillel, in association with rabbis representing other schools and interests. In the ensuing struggle for power, the patriarch managed to concentrate all communal authority in his office. The dominating role of the patriarchate reached its zenith in the days of his son and successor, Judah the Prince, whose reign (*c.* 175–*c.* 220) marked the climax of this period of rabbinic activity, otherwise known as the "age of the *tannaim*" (teachers). Armed with wealth, Roman backing, and dynastic legitimacy (which the patriarch now traced to the house of David), Judah sought to standardize Jewish practice through a corpus of legal norms that would reflect recognized views of the rabbinate on every aspect of life. The Mishna (collection of rabbinic law) that soon emerged became the primary source of reference in all rabbinic schools and constituted the core around which the Talmud (commentary on Mishna, literally "teaching") was later compiled. It thus remains the best single introduction to the complex of rabbinic values and practices as they evolved in Roman Palestine.

The Mishna

The making of the Mishna. Although the promulgation of an official corpus represented a break with rabbinic precedent, Judah's Mishna did have antecedents. During the 1st and 2nd centuries CE, rabbinic schools had compiled for their own reference collections in which the results of their exegesis and application of Scripture to problematic situations (Midrash, "investigation" or "interpretation"; plural Midrashim) had been recorded in terse legal form. By 200 CE several such compilations were circulating in Jewish schools and were being utilized by judges. While adhering to the structural form of these earlier collections, Judah compiled a new one in which universally accepted views were recorded alongside those still in dispute, thereby largely reducing the margin for individual discretion in the interpretation of the law. Although his action aroused opposition, and some rabbis continued to invoke their own collections, the authority of his office and the obvious advantages of a unified system of law soon outweighed centrifugal tendencies, and his Mishna attained quasi-canonical status, becoming known as "The Mishna" or "Our Mishna." For all its clarity and comprehensiveness, its phraseology was often obscure or too terse to satisfy all needs, and a companion known as the Tosefta ("Additions") was compiled shortly thereafter in which omitted traditions and explanatory notes were recorded. Since, however, neither compilation elucidated the processes by which their decisions had been elicited, various authorities set about collecting the midrashic discussions of their schools and recording them in the order of the verses of Scripture. During the 3rd and 4th centuries the tannatic Midrashim on the Pentateuch were compiled and introduced as school texts.

Fundamentally legal in character, this literature was designed to regulate every aspect of life—the six divisions of the Mishna on agriculture, festivals, family life, civil law, sacrificial and dietary laws, and purity encompass virtually every area of Jewish experience—and, accordingly, also recorded the principal Pharisaic and rabbinic definitions and goals of the religious life. One tract of the Mishna, *Avot* ("Sayings of the Fathers"), treated the meaning and posture of a life according to Torah, while other passages made reference to the mystical studies into which only the most advanced and religiously worthy were initiated; *e.g.*, the activities of the *Merkava*, or divine "Chariot," and the doctrines of creation (see below, *Jewish mysticism*). The rabbinic program of a life dedicated to study and fulfillment of the will of God was thus a graded structure in which the canons of morality and piety were attainable on various levels, from the popular and practical to the esoteric and metaphysical. Innumerable sermons and homilies preserved in the midrashic collections, liturgical compositions for daily and festival services, and mystical tracts circulated among initiates all testify to the deep spirituality that informed rabbinic Judaism.

THE AGE OF THE AMORAIM: THE MAKING OF THE TALMUDS (3RD–6TH CENTURIES)

Palestine (c. 220–c. 400). The promulgation of the Mishna initiated the period of the *amoraim* (lecturers or interpreters), those teachers who made the Mishna the basic text of legal exegesis. The curriculum now centred on the elucidation of the text of the standard compilation, harmonization of its decisions with extra-Mishnaic traditions recorded in other collections, and the application of its principles to new situations. The records of these amoraic studies have been preserved in the form of two running commentaries on the Mishna known as the Palestinian (or Jerusalem) Talmud ("Teaching") and the Babylonian Talmud, reflecting the study and legislation of the academies of the two principal centres of Jewish concentration in the Roman and Persian empires of that time. (Talmud is also the comprehensive term for the whole collections, Palestinian and Babylonian, containing Mishna, commentaries, and other matter. See below, *The literature of Judaism*)

The rabbinic educational system

The principal agencies mediating the rabbinic way of life and literature to the masses were the schools, ranging from the primary school to the advanced "house of study" and more formal academy (yeshiva), the synagogue, and the Jewish courts, which not only adjudicated litigations but also decided on ritual problems. Primary schools had long been available in the villages and cities of Palestine, and tannaitic law made education of male children a religious duty. Introduced at the age of five or six to Scripture, the student advanced at the age of 10 to Mishna and finally in mid-adolescence to Talmud or the processes of legal reasoning. Regular reading of Scripture in the synagogue on Mondays, Thursdays, the Sabbaths, and festivals, coupled with concurrent translations into the Aramaic vernacular and frequent sermons, provided for lifelong instruction in the literature and the values elicited from it. The amoraic emphasis on the moral and spiritual aims of Scripture and its ritual is reflected in their midrashic collections, which are predominantly homiletical (sermonic) rather than legal in content.

An amoraic sermon conceded that of every thousand beginners in primary school only one would be expected to continue as far as Talmud. In the 4th century, however, there were enough advanced students to warrant academies in Lydda, Caesarea, Sepphoris, and Tiberias (in Palestine), where leading scholars trained disciples for communal service as teachers and judges. In Caesarea, the principal port and seat of the Roman administration of Palestine, where pagans, Christians, and Samaritans maintained renowned cultural institutions, the Jews, too, established an academy that was singularly free of patriarchal control. The outstanding rabbinic scholar there, Abbahu (*c.* 279–320), wielded great influence with the Roman authorities and, because he combined learning with personal wealth and political power, attracted some of the most gifted students of the day to the city. In *c.* 350 the studies and decisions of the authorities in Caesarea were compiled as a tract on the civil law of the Mishna. Half a century later, the academy of Tiberias issued a similar collection on other tracts of the Mishna, and this compilation, in conjunction with the Caesarean material, constituted the Palestinian Talmud.

The role of the patriarchs

Despite increasing tensions between some rabbinic circles and the patriarch, his office was the agency providing a basic unity to the Jews of the Roman Empire. Officially recognized as a Roman prefect, a government official, the patriarch at the same time delegated apostles to Jewish communities to inform them of the Jewish calendar and of other decisions of general concern and to collect an annual tax of a half shekel paid by male Jews for his treasury. As titular head of the Jewish community of the mother country and as a vestigial heir of the Davidic monarchy, the patriarch was a reminder of a glorious past and of a hope for a brighter future. How enduring these hopes were may be seen from the efforts to gain permission to rebuild the Temple in Jerusalem. Although the emperor Julian (reigned 361–363) actually authorized the reconstruction, the project came to naught as a consequence of a disastrous fire on the sacred site and the subsequent death of the Emperor.

The adoption of Christianity as the religion of the empire had no direct effect on the religious freedom of the Jews; *i.e.*, on their freedom to worship and observe their life rules. The ever-mounting hostility between the two religions, however, resulted in severe curtailment of Jewish disciplinary rights over their coreligionists, interference in the collection of patriarchal taxes, restriction of the right to build synagogues, and, finally, upon the death of the patriarch Gamaliel VI in *c.* 425, the abolition of the patriarchate and the diversion of the Jewish tax to the imperial treasury. Though Mediterranean Jewry was now fragmented into disjointed communities and synagogues, the principles of the regulation of the Jewish calendar had been committed to writing *c.* 359 by the patriarch Hillel II, and this, coupled with the widespread presence of rabbis, ensured continuity of Jewish adherence. Even the emperor Justinian's (reigned 527–565) restrictions on synagogal worship and preaching apparently had no devastating effect. A new genre of liturgical poetry, combining ecstatic prayer with didactic motifs, developed in this period of political decline and won acceptance in synagogues in Asia Minor as well as beyond the Euphrates.

Babylonia (200–650). In the increasingly unfriendly climate of Christendom, Jews drew consolation in the knowledge that in nearby Babylonia (then under Persian rule) a vast population of Jews continued to live under a network of effective and autonomous Jewish institutions and officialdom. Steadily worsening conditions in Palestine had drawn many Jews to Persian domains, where economic opportunities and the Jewish communal structure enabled them to gain a better livelihood while living in accordance with their ancestral traditions. To regulate internal Jewish affairs and ensure the steady flow of taxes, the Parthian, or Arsacid, rulers (247 BCE–224 CE) had appointed *c.* 100 CE an exilarch, or "head of the [Jews in] exile"—who claimed more direct Davidic descent than the Palestinian patriarch—to rule over the Jews as a quasi-prince. In *c.* 220 two Babylonian disciples of Judah the Prince, Abba Arika and Samuel bar Abba, began to propagate the Mishna and related tannaitic literature as the yardsticks of normative practice. As heads of the academies at Sura and Nehardea, respectively, Abba and Samuel cultivated a native Babylonian rabbinate, which increasingly provided the manpower for local Jewish courts and other communal services. While the usual tensions between temporal and religious arms frequently erupted in Babylonia, too, the symbiosis of exilarchate and rabbinate endured uninterruptedly until the middle of the 11th century.

Paradoxically, Babylonian rabbinism derived its ideological strength from its fundamentally unoriginal character. As a transplant of Palestinian Judaism it claimed historic legitimacy to the Sāsānid rulers (224–651), who protected Jewish practices against interference from fanatical Magian priests, and to native Jewish officials, who argued for the validity of indigenous Babylonian deviations from Palestinian norms. But ultimately the historic importance of this transplantation lay in Babylonia's serving as the proving ground for the adaptability of Palestinian Judaism to a Diaspora situation. Legal and theological adaptations generated by needs of the new locale and times inevitably effected changes in the religious tradition. The laws of agriculture, purity, and sacrifices all of necessity fell into disuse. The values embodied in these laws, however, and the core of the legal–theological system—ranging from doctrinal faith in the revelation and election of Israel, to the requirement that the individual live by the canons of Jewish civil and family law, and the establishment of a network of communal institutions modelled on those of the mother country—remained intact, thereby ensuring a basic continuity and uniformity to rabbinically oriented communities everywhere. The real contribution of the Babylonian rabbinate to Jewish religion lay, accordingly, in its demonstration of how Palestinian Judaism was to be implemented on Gentile soil. Since historic circumstances made Babylonia the mediator of this tradition to all Jewish communities in the High Middle Ages (9th–12th centuries), the Babylonian version of Jewish religion became synonymous with normative Judaism and the measure of Judaic authenticity everywhere.

The model for Diaspora Judaism

"The law of the [Gentile] government is binding," the principle formulated by Samuel, head of the academy at Nehardea (died 254), summarizes the essential novelty in rabbinic reorientation to life on foreign soil. Whereas Palestinian rabbis had perforce to comply with imperial decrees of taxation de facto—and this was all that Samuel had in mind—Babylonian teachers now rationalized the legitimacy of governmental authority in this respect de jure and thus enjoined upon the Jews political quietism and submissiveness as part of their religious theory. In all other areas of civil law, the Jews were instructed by their rabbis to bring their litigations to Jewish courts and thus to conduct their businesses as well as their family lives by rabbinic law.

While the rabbis could obviously more effectively impose their discipline in matters of public law than in private religious practice, the density of the Jewish population in many areas of Parthia (northeastern Iran) and Babylonia facilitated the application of moral and disciplinary pressures. The most effective vehicle for the dissemination of their teachings was the academies, of which those of Sura and Pumbedita remained preeminent, where judges and communal teachers were trained. Frequent public lectures in the synagogues of the academies on Sabbaths and festivals were capped by public *kalla* (study-course) assemblies for alumni of the schools during the two months, Adar (February–March) and Elul (August–September), when the lull in agricultural work freed many to attend semiannual refresher instruction. These meetings were followed by regular popular lectures during the festival seasons that soon followed. Thus, while rabbis constituted a distinct class within the community, their efforts were oriented toward making as much of the community as possible members of an elite of learning and religious scrupulosity. The harmonious relations that obtained with but few interruptions over the centuries between the Sāsānian rulers and their Jewish subjects gave the Jewish population the air of a quasi-state, which the Jewish leadership frequently extolled as superior to the Jewish community of Palestine.

The dissemination of the Palestinian Talmud probably stimulated the Babylonians to follow suit by collecting and arranging in similar fashion the records of study and decisions of their own academies and courts. The Babylonian Talmud, which apparently underwent several stages of redaction (*c.* 500–650) on the basis of the proto-Talmuds—the early collections of commentaries on the Mishna—used in the academies, accordingly became the standard of reference for judicial precedent and theological doctrine for all of Babylonian Jewry and all those communities under its influence. As had been the case with the Mishna, the redaction of the Babylonian Talmud was later designated by authorities as marking the end of a period in Jewish history, and the scholars who put the finishing stylistic touches, known as *savora'im* ("explicators"), were classified as a transitional stage between the *amoraim* and *geonim* (see below).

Redaction of the Babylonian Talmud, *c.* 500–650

The enduring vigour of Jewish faith throughout these centuries is graphically demonstrated by the missionary activity of Jews throughout the ancient Middle East, especially in the Arabian Peninsula. Proud Jewish tribes living in close proximity to each other in the vicinity of Yathrib (later Medina, Muḥammad's home city), engaged in agriculture and commerce and also in proclaiming the superiority of their monotheistic ethos and eschatology (doctrine of last things). In Yemen (southwestern Arabia) the last of the Ḥimyarite rulers (reigned from *c.* 2nd century CE), Dhu Nuwas, proclaimed himself a Jew and finally suffered defeat (*c.* 525) as a consequence of Christian influence on the Abyssinian armies. Jewish missionaries, however, continued to compete with Christian missionaries and thus helped lay the groundwork for the birth of an indigenous Arabic monotheism—Islām—that was to alter the course of world history.

THE AGE OF THE GEONIM (C. 640–1038)

Triumph of the Babylonian rabbinate. The lightning conquests in the Middle East, North Africa, and the Iberian Peninsula by the armies of Islām (7th–8th centuries) provided the environmental framework for the basically uniform (*i.e.*, Babylonian) character of medieval Judaism. As a "people of the Book" (*i.e.*, of the Bible), the Jews were permitted by the Muslims to live under the same autonomous structure that had developed under Arsacid and Sāsānian rule. The heads of the two principal academies were now formally recognized by the exilarch, and through him by the Muslim caliphate (religiopolitical rulers), as the official arbiters of all questions of religious law and as the religious heads of all Jewish communities that came under Muslim sway. Known as *geonim* (plural of *gaon*, "excellency"), and conducting high courts manned by scholars assigned graded ranks, they drew their financial support from Jewish communities assigned to them by the exilarch. Religious questions and contributions were solicited from all Jewish communities, and these along with formal gaonic replies (*responsa*) were regularly publicized at the semiannual *kalla* convocations. Under the strong leadership of Yehudai, *gaon* of Sura (presided 760–763), the Babylonian rabbinate exerted vigorous efforts to replace Palestinian usage wherever it was still in vogue, including the study of Palestinian amoraic legal literature, by Babylonian practice and texts, thus making the Babylonia Talmud the unrivalled standard of Jewish norms everywhere. The success of this campaign is evidenced by the fact that the term Talmud, when unqualified, has ever since meant the Babylonian Talmud. Indeed, even in Palestine the Babylonian corpus displaced its older rival and caused the study of Palestinian Talmudic literature to be confined to circles of legal specialists.

Antirabbinic reactions. The firm, and on occasion oppressive, tactics of exilarchs and *geonim* generated antirabbinic reactions, especially in outlying areas where enforcement was difficult, in the form of sectarian and messianic revolts. Inspired in part by ancient Palestinian sectarian doctrines and in part by Muslim usage, the sects were by and large quickly and forcefully suppressed. In the 9th century, however, a moderate group under the leadership of Anan ben David, a disaffected member of the exilarchic family, successfully organized a dissident movement that soon developed into a formidable challenger of Rabbinite (a term first used for the Talmudic adherents by the dissidents) supremacy. Known as Karaites (Scripturalists), the new sect advocated a threefold program of (1) rejection of rabbinic law as a human fabrication and therefore an unwarranted, unauthoritative addition to Scripture, (2) a return to Palestine to hasten the messianic redemption, and (3) a re-examination of Scripture to retrieve authentic law and doctrine. Under the leadership of Daniel al-Qumisi (*c.* 850?), a Karaite settlement prospered in the Holy Land, from which it spread as far as northwestern Africa and Christian Spain. A barrage of Karaite treatises arguing new views of scriptural exegesis stimulated renewed study of the Bible and Hebrew language in Rabbinite circles as well. The most momentous consequence of these new studies was the invention of several systems of vocalization for the text of the Hebrew Bible (Christian Old Testament) in Babylonia and Tiberias in the 9th and 10th centuries. The annotation of the Masoretic (traditional, or authorized) text of the Bible with vocalic, musical, and grammatical accents in the Tiberian schools of the 10th-century scholars Ben Naftali and Ben Asher fixed the Masoretic text permanently and through it the morphology (basic form and structure) of the Hebrew language for Karaites as well as Rabbinites.

The Karaite movement

In the face of sectarian challenges, the *geonim* intensified their efforts against any deviation from Rabbinite norms and began to issue handbooks of Jewish law that set forth in concise and unequivocal terms the standards for correct practice. A number of these codes, notably the *Halakhot gedolot* ("Great Laws"), *Siddur Rav Amram Gaon* (on liturgical practice), and *She*ʾ*eltot* ("Disquisitions") by Aḥa of Shabḥa (*c.* 680–*c.* 752), attained authoritative status in local schools and further helped give a unitary stamp to medieval Judaism.

Secular culture and philosophy

The *geonim,* however, were powerless to halt several social developments in the 9th century that progressively undermined their hold even over Rabbinite communities. A renascence of Greek philosophy and sciences in Arabic translation, coupled with the progressive urbanization of the upper classes of all religioethnic groups in the centres of political, commercial, and cultural activity, generated a new intelligentsia that cut across religioethnic lines. Widespread skepticism in basic doctrines of faith such as creation, revelation, and retribution was most poignantly represented by latitudinarianism (the tendency to be flexible and tolerant about deviations from orthodox beliefs and doctrines) and by antinomian (anti-Mosaic-law) Gnostic groups that negated divine providence and omniscience. Ḥiwi al-Balkhi, a 9th-century skeptical Jewish pamphleteer, scandalized the faithful by an open attack on the morality of Scripture and by an expurgated edition of the Bible for schools that omitted "offensive" materials (*e.g.*, alleged stories of God acting dishonestly). A mystifying Hebrew tract entitled *Sefer yetzira* ("Book of Creation") posited in terse and enigmatic epigrams a novel theory of creation that betrayed unmistakable Neoplatonic influence. Karaites joined philosophically oriented intellectuals in heaping scorn on popular Rabbinite customs that smacked of superstition and, above all, on Talmudic homilies that referred to God in anthropomorphic terms.

Gaonic difficulties were compounded by the rise in North Africa and Spain of populous and wealthy Jewish communities that, thanks to the development of their own local schools and native talent, ignored the Babylonian academies or favoured one over the other with religious queries and, in consequence, with financial contributions. To the delight of dissidents and the chagrin of the faithful, competition between the Babylonian academies turned to internecine hostility. Occasional revolts against exilarchic taxation and administration in outlying areas of Persia had to be quelled with armed force. The Palestinian Rabbinites had revived their own academies, and their presidents now not only appealed for support in other Diaspora lands but challenged the authority of the Babylonians to serve as final arbiters on such matters of public import as the regulation of the calendar. By 900 the Rabbinite community of Babylonia was in a state of chaos and dissolution.

The gaonate of Saʿadia ben Joseph. In a bold effort to restore discipline and respect for the gaonate, an able exilarch, David ben Zakkai (916/917–940), bypassed the families from whom the *geonim* had traditionally been selected and in 928 appointed Saʿadia ben Joseph al-Fayyumi to head the academy of Sura. Of Egyptian birth, Saʿadia had gained wide acclaim for his scholarly retorts to Karaites, heretics, and Palestinian Rabbinites. Politically, Saʿadia's brief presidency was a fiasco and aggravated the chaos by a communal civil war. His gaonate, however, gave an official stamp to his many works, which responded to the ideological challenges to Rabbinism by restating traditional Judaism in intellectually cogent terms. Saʿadia thus became the pioneer of a Judeo-Arabic culture that was to come to full flower in Andalusian Spain a century later (see below *Sefardic developments*). His translation of the Bible into Arabic and his Arabic commentaries on Scripture made the rabbinic understanding of the Bible accessible to masses of Jews. His poetic compositions for liturgical use provided the stimulus for the revival of Hebrew poetry. Above all, his rationalist commentary on the puzzling "Book of Creation" and his brilliant philosophic treatise on Jewish faith, *Beliefs and Opinions,* synthesized Torah (the divine law in the Five Books of Moses and the rabbinic understanding of this revelation) and "Greek wisdom" in accordance with the dominant Muslim philosophical school of Kalām and thus made Judaism philosophically respectable and the study of philosophy a religiously acceptable pursuit.

Far from tightening the gaonic hold over the Jewish communities of the Arabic world, Saʿadia's works actually provided the wherewithal for ever-greater intellectual and religious self-sufficiency. While economic, political, and military upheavals progressively weakened all institutional fabrics in the Middle East, concurrent prosperity and consolidation in the West stimulated the maturation of indigenous leadership in Egypt, al-Qayrawān (Kairouan; in present-day Tunisia), and Muslim Spain. To be sure, able *geonim* such as Sherira and his son Hai exercised enormous influence over the Judeo-Arabic world through hundreds of legal *responsa* issued in the course of their successive terms (968–1038) at Pumbedita. Circumstances beyond anyone's control, however, were bringing the curtain down on the effectiveness of exilarchate and gaonate. But by 1038, the year of Hai's death, the consequences of four centuries of gaonic activity had become indelible: the Babylonian Talmud had become the agent of basic Jewish uniformity; the synthesis of philosophy and tradition had become the hallmark of the Jewish intelligentsia; and the Hebrew classics of the past had become the texts of study in Jewish schools everywhere.

Gaonates of Sherira and Hai

MEDIEVAL EUROPEAN JUDAISM (950–1750)

The two major branches. Despite the fundamental uniformity of medieval Jewish culture, the cultural–political divisions within the Mediterranean basin, in which Arabic-Muslim and Latin-Christian civilizations coexisted as discrete and self-contained societies, shaped the character of the Jewish subculture of the area. Two major branches of rabbinic civilization developed in Europe, the Ashkenazic, or Franco-German, and the Sefardic, or Andalusian-Spanish. Distinguished most conspicuously by their varying pronunciation of Hebrew, the numerous differences between them in religious orientation and practice derived, in the first instance, from the geographical fountainheads of their culture—the Ashkenazim (plural of Ashkenazi) tracing their cultural filiation to Italy and Palestine and the Sefardim (plural of Sefardi) to Babylonia—and from the influences of their respective immediate milieus. While the Jews of Christian Europe wrote for internal use almost exclusively in Hebrew, those of Muslim areas regularly employed Arabic for prose works and Hebrew for poetic composition. Whereas the literature of Jews in Latin areas was overwhelmingly religious in content, that of the other branch was well endowed with secular poetry and scientific works inspired by the cultural tastes of the Arabic literati. Most significantly, the two forms of European Judaism differed in their approaches to the identical rabbinic base that both had inherited from the East and in their radically different attitudes to Gentile culture and politics.

Sefardic developments. In Muslim Spain, Jews frequently served the government in official capacities and, therefore, not only took an active interest in political affairs but also engaged in considerable social and intellectual intercourse with influential circles of the Muslim population. Since the support of letters and scholarship was part of state policy in Muslim Spain, and since Muslim savants traced the source of Muslim power to the vitality of the Arabic language, scripture, and poetry, Jews looked at Arabic culture with undisguised admiration and unabashedly attempted to adapt themselves to its canons of scholarship and good taste. The hallmark of the cultured Jew accordingly became a polished command of Arabic style and the ability to display the beauty of his own heritage through a philological mastery of the text of the Hebrew Bible and through the composition of new Hebrew verse, now set to an alien Arabic metre. Since Arabic philosophers and scientists promulgated syntheses of Greek philosophy with the revelation to Muḥammad, rationalist study of the Jewish classics and defense of rabbinic faith in philosophic terms became dominant motifs in the Andalusian Jewish schools (in southern Spain).

The Golden Age of Jewish Arabic and Hebrew culture

The atmosphere generated a fever of literary creativity in classical Jewish disciplines as well as in the sciences cultivated by the Arabs that has gained for the period the title of "the Golden Age of Hebrew literature" (*c.* 1000–1148). What distinguished the Jewish culture of this age was not only the supreme literary merit of its Hebrew poetry, the new spirit of relatively free and rationalist examination of hallowed texts and doctrines, and the extension of Jewish cultural perspectives to totally new horizons—mathematics, astronomy, medicine, philosophy, political theory, aesthetics, belles-lettres—but also the frequent overlapping of the Sefardic religious leadership with the new Jewish courtier class. The unprecedented heights which the latter attained—Ḥisdai ibn Shaprut as counsellor to the caliphs of Córdoba, the Ibn Nagrelas as viziers of Granada, the Ibn Ezras, Ibn Megashs, and Ibn Albalias as high officials in Granada and Seville—and the distinctions of these men and of their protégés in Jewish and worldly letters restored the ancient integration of culture and practical life and generated a neoclassicism ("classicism" here meaning biblicism) that expressed the identification of the Jewish elite with the biblical age of Jewish power and artistic creativity. The effort to recapture the vitality and beauty of biblical poetry stimulated comparative philological and fresh exegetical research that yielded new insights into the morphology of the Hebrew language and into the historical soil of biblical prophecy. Judah ibn Ḥayyuj and Abū al-Walīd Marwān ibn Janāḥ produced manuals on biblical grammar that applied the results of Arabic philology to their own tongue and that have, accordingly, provided the principles of Hebrew grammatical study down to modern times. The anticipations of modern higher biblical criticism by Judah ibn Balaʿam and Moses ibn Gikatilla (flourished 11th century) were popularized in Hebrew a few generations later by Abraham ibn Ezra. In the revival of Hebrew poetry, liturgical as well as secular, that translated the new preoccupation with language and beauty into art, Andalusian Jewry saw its greatest achievements. Solomon ibn Gabirol, Moses ibn Ezra, and Judah ha-Levi were but the acknowledged supreme geniuses of a form of expression that became a passion with thousands the length and breadth of Spain. But by far the most enduring consequence of the new temper was their redefinition of religious faith in the light of Greco-Arabic philosophical theories. Solomon ibn Gabirol's exposition of faith in Neoplatonic terms, Abraham ibn Daud's defense of Rabbinism by Aristotelian categories, Judah ha-Levi's attack on philosophy as religiously bankrupt, and Moses Maimonides' epoch-making synthesis of Judaism and medieval Aristotelianism fixed philosophic inquiry as an enduring subject on the agenda of rabbinic concerns. A new class of philosophers that emerged in the 13th century and sponsored the translation of Arabic literature into Hebrew and of Hebrew and Arabic literature into Latin brought Jews and their thought into the mainstream of Western philosophy and gained for them the position of middlemen of culture between East and West.

Influence of Greco-Arabic philosophy

The salient trends of Sefardic Judaism did not imply relegation of the rabbinic class to a second place. Rather they shaped a fresh approach to rabbinic texts that paralleled in many respects those adopted in biblical exegesis. Strict adherence to consistency, systematization, and philological exactitude yielded new codes that often diverged from gaonic judgments. A digest of Talmudic law by Isaac Alfasi placed the Sefardic rabbinate on a self-reliant footing and epitomized its ideal of getting at the essentials of Talmudic law by sidestepping contingent discussions. In this area, too, it was Moses Maimonides who through his code of Jewish law, *Mishne Torah,* brought the Sefardic principles of comprehensiveness, lucidity, and logical arrangement to their apex. Written in Mishnaic Hebrew, the work remains to this day the only comprehensive treatment of all of Jewish law, including those fields that are not applicable in the Diaspora (agriculture, purity, sacrifices, Temple procedure).

With Maimonides, however, the pure Sefardic tradition came to an end, for the Almohad (Berber Muslim reformers) invasion of Spain in 1147–48 wiped out the Jewish communities of Andalusia and drove thousands either to northern Spain and Provence or, as in the case of Maimonides' family, to North Africa and Egypt. Sefardic Jewry suddenly encountered a discrete, mature, Jewish culture that for centuries had been developing independently and along quite different lines.

Ashkenazic developments. The spokesmen of Ashkenazic Jewry, into whose communities the Sefardim had been thrust by political events, regarded their own heritage and the Christian world in which they lived from a perspective shaped exclusively by rabbinic categories. From the world of the Talmud and Midrash they drew their school texts and the values that determined their judgments. Sensing no intellectual challenge in Christian faith, which they regarded with thinly concealed contempt, they constituted for the most part a merchant class that lived in urban centres under the protection of ecclesiastical and temporal rulers but under their own complex of laws and institutions. Except for mercantile relations, Christian society was closed to them, thanks largely to age-old ecclesiastical prohibitions forbidding all social intercourse with them. With the Arab conquest and the rise of the Carolingians (the 8th–10th-century dynasty that ruled France and Germany), the 12-decade interlude of suppression by the Visigoths (589–711) came to an end, and the Roman precedent of toleration and autonomy again became the rule. Merchants and rabbis moved from Italy to France and the Rhineland and infused new energies into the Jewish communities there. A native religious leadership began to emerge at the very time that Andalusian Jewry was entering its Golden Age. The bloody upheavals of the First Crusade (1096–99) in the communities of the Rhineland, although unleashing a tide of hatred, periodic violence, and progressive restrictions on Jewish activities, struck Jewish communities that had attained sufficient resilience to reestablish their communal institutions shortly afterward and continue the cultivation of their deeply ingrained traditions.

Ashkenazic cultural and religious patterns

By 1150 Ashkenazic Jewry had generated a culture pattern of its own with an indigenous literature that ranged from the popular homily to the esoteric tract on the nature of the divine glory. Study of the Bible and Talmud was oriented toward a mystical pietism in which prayer and contemplation of the secrets embedded in the liturgy were to lead to religious experience. Significantly, the fathers of the Ashkenazic tradition were remembered as liturgical poets and initiates into divine mysteries, and the early codes of the Franco-German schools were heavily weighted with discussions of liturgical usage. After the Second Crusade (1147–49), the German Jewish mystics (also called Ḥasidim, or pietists) placed heavy emphasis on the merits of asceticism, martyrdom and lifelong disciplines of penitence, thus adapting to Jewish idiom the features of saintliness celebrated in the universe of discourse of which they were a part. For the masses of Jews the cultural fare consisted principally of biblical tales and instruction, as interpreted by rabbinic Midrash, the lives of scholars and saints, and liturgical poetry reaffirming the election of Israel and faith in messianic redemption. The chief vehicle of popular instruction consisted of anthologies from the Rabbinic writings and commentaries on Scripture, of which the most popular was that of Rabbi Solomon ben Issac of Troyes, known as Rashi, the acronym formed from the initials of his name in Hebrew. For the more advanced student, Rashi composed a succinct commentary on the Talmud that, unmatched for compact thoroughness and lucidity, achieved an authority approaching that of the text itself.

As living sources of law and values, the Bible and Talmud had an impact that was apparent in communal decision and in the bearing of the leadership at home, in the marketplace, and in the synagogue. Taking their cue from Talmudic precedent and from Christian ecclesiastical procedures of their own times, the Ashkenazic rabbis occasionally gathered in regional synods to enact legislation on problems of a general nature for which there was no adequate precedent in the literature. Among the most enduring of these measures were the prohibition of bigamy and arbitrary divorce and severe economic penalties for abandonment of wives. Of far more immediate concern to the average Jew were the circumvention of Talmudic prohibitions against usury, relaxation of prohibitions regarding traffic with Gentiles in wines, and adoption of severe disciplinary measures, such as excommunication, against informers or those appealing, in cases involving Jews, to the Gentile authorities.

The rise and spread of Kabbala

A new religious trend began in Provence (a province of southeastern France) in the 13th century with the introduction into the Talmudic academies of a novel form of mystical study known as Kabbala (literally, "tradition"), which soon spread to northern Spain. Expressing Gnostic-type doctrines in rabbinic guise, the devotees of Kabbala devised an esoteric vocabulary that reinterpreted the Bible and rabbinic law as allegories of the various modes in which God is manifested in a spiritual universe, access to which was reserved for initiates. The most renowned literary product of this new circle was the *Zohar* ("The Book of Splendour"), a vast mystical commentary on the Pentateuch by Moses de León (*c.* 1275), which with later additions became the Bible of Jewish mystics everywhere. Although some of the theological notions of the Kabbalists deviated from basic postulates of Jewish monotheism, the insistence of the mystics on unflagging ritual orthodoxy and on a nominal acceptance of the biblical text as divine revelation helped them avert the suspicions aroused by Jewish Aristotelians and Averroists (followers of the 12th-century Arabic Aristotelian philosopher Averroës) and, in time, even won for them the status of a rabbinic elite. Indeed, some of the mystics lent their support to the antiphilosophic campaign that began in Montpellier, in southern France, *c.* 1200 and condemned the study of philosophy as generating skepticism, latitudinarianism, and disrespect for traditional literature. (For a fuller discussion of Kabbala see below, *Jewish mysticism*.)

Conflicts, disasters, and new movements. Basically, the conflict between "fundamentalist" and philosopher in Provence and northern Spain represented a clash between two mature Jewish subcultures of diverse geographic origins, the Sefardic and Ashkenazic, each of which had in the course of centuries developed different esoteric doctrines to transcend the legalistic formalism and confining

Jews, longing for a return to the Holy Land, point to a visionary Jerusalem, which is depicted in the Gothic style of Christian church architecture. The Jews are shown with the pointed hats they were required to wear to distinguish them from Christians, and are represented with birds' heads, since they felt it was sacrilegious to depict the human form in sacred objects. Illustration from the Birds' Head Haggada, an illuminated manuscript from southern Germany, c. 1300.

By courtesy of the Israel Museum, Jerusalem

dogmas of normative Judaism. Both forms of speculation sought salvation for exceptional individuals through knowledge and thus provided an immediate substitute for messianic deliverance from exile and servitude. Each group charged the other with distortion of tradition, and each issued apologias (defenses or justifications) and excommunications characteristic of medieval doctrinal controversy. While the rifts within communities attained bitter proportions, the common threat posed by ecclesiastical attacks on the Talmud in public disputations and by the expulsion of the Jews from France in 1306 prevented open rupture or resolution of the conflict. Ever since that time, two strands of orthodoxy representing the two forms of medieval metaphysical speculation have lived side by side in an uneasy truce.

Most rabbinic circles of the 14th and 15th centuries displayed a progressive dogmatism and insistence on uniformity of practice. The great legal code of Jacob ben Asher of Toledo, *Arbaʿa ṭurim* (*c.* 1335; "Four Rows"), which sought to level differences in usage between Ashkenazim and Sefardim, bespoke the dominant trend of the rabbinate. The increasing hardening of ideological lines, however, did not eliminate independent thinking. Gersonides (Levi ben Gershom) gave Jewish Aristotelianism a new and comprehensive formulation, while Isaac Albalag propounded an Averroist (rationalistic) interpretation of the Bible predicated on a theory of double truth (of reason and revelation). In Muslim areas, the Maimonidean regimen of philosophic contemplation was extended by Maimonides' son Abraham to a quest for pietist ecstasy that betrayed many features of Ṣūfism (Islāmic mysticism).

Anti-Jewish riots and massacres of 1391 and a wave of apostasy in the wake of the disputation of Tortosa (1411–14)—which ended with a papal bull forbidding Talmud study, compelling attendance at Christian sermons, and other onerous measures—struck catastrophic blows in the Spanish communities and fed the anti-intellectualism of the rabbinate. Hasdai Crescas, while conceding the philosophic untenability of traditional belief in freedom of the will, launched a scathing attack on Aristotelian approaches to religion, and his disciple Joseph Albo issued a compendium on dogma that reaffirmed the traditional postulates of divine creation, revelation, and retribution as axioms of Judaism. But these reassertions of traditional faith could not overcome the ideological and social fragmentation that had split the Spanish communities into congealed strata that were often in open conflict with each other. Widespread marranism (ostensible conversion to Christianity) polarized the community and left deposits of bitterness that extended to those returning to the fold. The expulsions from Spain (1492) and Portugal (1497 and 1506) dealt the final blow and drove the escaping leadership into intensified pursuits of mystical escape from, and rationalization of, the endless calamities that befell their flocks. In Italy and the Ottoman Empire (Asia Minor, northeastern Africa, and southeastern Europe), the two principal centres of refuge for the exiles of the Iberian Peninsula, legalistic Kabbalism, which insisted on strict observance of the law as precondition of mystical practice and study, became the dominant spirit of a rabbinic leadership that in the face of terribly adverse circumstances continued to produce works of encyclopaedic proportions and staggering erudition in every field of Jewish learning.

Adversity and response: mysticism, messianism, rationalism

Inspired by the Jewish tradition that the messianic era—when the messiah would come to bring in the rule of God—would be preceded by horrendous catastrophes, a group of single-minded rabbis established a community in Ẕefat (Safed), Palestine, where in anticipation of the new dawn all of life was to be conducted on principles of saintliness and mystical contemplation. Under the leadership of one Jacob Berab, the ancient practice of ordination was reinstituted in 1538 to form the nucleus of a revived Sanhedrin so as to administer ritual procedures requiring ordained authorities. While the effort failed because of rabbinic opposition, it reflected a widespread temper and further fanned messianic hopes sparked shortly before by the campaigns of tragic consequences by David Reubeni and Solomon Molkho in Italy, which ended in their being burned at the stake by the Christian authorities. In Ẕefat itself Kabbalism soon entered a new phase under the inspiration of Isaac Luria and Ḥayyim Vital, who confided to their disciples that the calamities of Israel were but a mirror of the captivity into which many sparks of the Godhead itself had fallen. Liturgical innovations and a novel mystical theology were formulated to redeem the imprisoned elements of divinity and thus restore creation to the harmony intended for it (see also below, *Jewish mysticism*).

That the Almighty himself was not quite omnipotent, at least with respect to the fate of his chosen people, was cautiously hinted in a Hebrew work of history (1550) by Solomon ibn Verga, who saw the Jewish problem as a sociopolitical one to which theological answers were futile. Such guarded rationalism was entertained by a number of courageous thinkers in 16th-century Italy, where, despite the policy of ghettoization (the segregation of the Jewish community in a restricted quarter) begun by Venice in 1516 and soon extended to all major Italian cities, the spirit of the Renaissance and the passion for historical criticism had captivated many Jews. Catholic scholars and prelates occasionally employed rabbis to instruct them in the Hebrew language and in the secrets of the Kabbala, which some Christians believed actually verified the postulates of their own faith. Contacts with Christian scholars in turn introduced Jews like Azariah dei Rossi, whose *Meʾor ʿenayim* ("Enlightenment of the Eyes") inaugurated critical textual study of rabbinical texts, to new bodies of literature that had been lost to the Jewish community, such as the works of Philo and Josephus (see above *Hellenistic Judaism*).

The Shabbetaian debacle

Such phenomena, however, were decidedly in the minority and contrary to the dominant trend. Dogmatic Kabbalism spread progressively and finally came to social expression in 1666 with the widespread acceptance of the views of the pseudo-messiah Shabbetai Tzevi (Sabbatai Zevi). Most of European and Ottoman Jewry was swept into a hysterical pitch in the belief that the end was now finally at hand. When the pseudo-messiah converted to Islām after being apprehended by the Ottoman government, mass despondency took the form of crypto-Shabbetaianism in which the apostasy of the messiah was explained as a form of voluntary crucifixion for the sake of the Jews. A witch-hunt on the part of traditionalists to uncover the cells of heresy unsettled Jewish communities everywhere by an emphasis on greater rigidity than before.

The following century (to *c.* 1750) was the darkest in the history of rabbinic Judaism. Scholarship reached an ebb of quality and popular religion a mechanical state such as Jews had never before experienced. The massacres and impoverishment of Polish Jewry after 1648 brought a pall over the growing eastern European centres of Jewish life. Antinomian eruptions of extreme Shabbetaians under the leadership of the self-proclaimed messiah and later Catholic convert Jacob Frank (1726–91) alarmed Gentile authorities almost as much as they did Jews. But the fossilization referred to above was only apparent. Beneath the surface many were restlessly searching for new avenues of faith, and the 18th century saw fresh responses that set the history of the Jews and of Judaism on new directions and spelled the beginnings of a new era. (G.D.C.)

Modern Judaism (c. 1750 to the present)

THE NEW SITUATION

The various criteria used to mark off dividing points in the history of the Jews and Judaism (see above *General observations*) are especially notable when it comes to setting a starting date for the modern period. Historians of thought put it in the late 17th century with the appearance of men, such as the philosopher Benedict de Spinoza, who ceased, in part or in toto, to believe in the inherited faith without at the same time ceasing to be Jews (*i.e.,* to consider themselves and be considered as Jews). Some Israeli scholars set it at about 1700 with the first stirrings of that new and continuing emigration from the Diaspora to the Holy Land that culminated in the mid-20th century in the creation of the State of Israel. Political and social historians set it in the mid- and late-18th-century processes that led to the American and French revolutions and to the results that flowed from these two epochal events, among them the emancipation of Jews from discriminatory and segregative laws and customs, the attainment of legal status as citizens, and the freedom of individual Jews to pursue careers appropriate to their talents. These varying approaches appear to have one thing in common—the view that these postmedieval forms of Jewish experience assume the end of the doctrine of the Exile, whereby Jews saw themselves as a people waiting out centuries of woe in alien lands until the moment of divine redemption. Jewish modernity for most scholars, then, is marked by the end of a passive waiting on the Messiah and the beginning of an active pursuit of personal or national fulfillment on this earth and preferably in one's lifetime.

Although the 18th century Haskala (Enlightenment) among the Ashkenazim of central and eastern Europe is often taken as the starting point of Jewish modernity, the process of Westernization had begun a good deal earlier among the Sefardim in western Europe and Italy. The Marranos who went to such communities as Amsterdam and Venice in the 17th century to declare themselves as Jews carried with them the Western education that they had acquired while living as Christians in the Iberian Peninsula, and the habits of criticism that had kept them from assimilating into the majority during their Marrano years and that some, such as Spinoza, a son of Marranos, used in analyzing all of the biblical tradition, including especially their own religion. In Italy there was an older Jewish community that had never been sealed off culturally from the influence of its environment; some of its figures were influenced by, and participated in, the main currents of the Renaissance (see above *Rabbinic Judaism*).

Increased contact with Western languages, manners, and modes of living came to the Ashkenazim only in the 18th century when new economic opportunities created such possibilities and needs. Jewish bankers and factors in various German principalities, army provisioners in most of the European countries, capitalists who were permitted to live in such places as Berlin because they opened new factories or were otherwise helpful to the expansion of the economy—all were in increasing contact with Gentile society, and most of them began to look upon the goal of their lives as the winning of full acceptance. Around this wealthy element there arose a number of intellectuals who agitated for the end of ghetto ways as the necessary preamble to the emancipation of the Jews.

THE HASKALA, OR ENLIGHTENMENT

The role of Moses Mendelssohn

In central Europe. By far the most outstanding figure of the 18th-century Jewish Enlightenment was the philosopher Moses Mendelssohn, who, while remaining a devoted adherent of Orthodox Judaism, turned away from the traditional Jewish preoccupation with the Talmud and its literature to the intellectual world of the European Enlightenment, of which he became the foremost Jewish representative. Mendelssohn did not attempt a philosophical defense of Judaism until pressed to do so by Christians who questioned how he could remain faithful to what they saw as an unenlightened religion. In his response, *Jerusalem,* published in 1783, Mendelssohn defended the validity of Judaism as the inherited faith of the Jews by defining it as revealed divine legislation and declared himself at the same time to be a believer in the universal religion of reason, of which Judaism was but one historical manifestation. Aware that he was accepted by Gentile society as an "exceptional Jew" who had embraced Western culture, Mendelssohn's message to his own community was to become Westerners, to seek out the culture of the Enlightenment. To that end he joined with a poet, Naphtali Herz (Hartwig) Wessely, in translating the Torah into German, combining Hebrew characters with modern German phonetics in an effort to displace Yiddish, and wrote a modern biblical commentary in Hebrew, the *Beʾur* ("Commentary"). Within a generation, Mendelssohn's Bible was to be found in almost every literate Jewish home in central Europe and had served to introduce its readers to German culture. Through his personal example and his life's work Mendelssohn made it possible for his fellow Jews to join the Western world without sacrificing their Judaism; he had convinced them that their intellectual processes were those of universal reason, with which Judaism accorded.

Educational reform

Mendelssohn's work was carried on by a group of Jewish intellectuals who had gathered around him in his lifetime, forming the nucleus of the Berlin Haskala, which was most active in the 20 years following their mentor's death. In the pages of their Hebrew-language periodical, *ha-meʾassef* ("The Collector"), they preached the virtues of secular culture and used Hebrew as a vehicle by which to introduce that culture. To achieve their goal of an enlightened Judaism, the leaders of the Berlin Haskala publicized the need for secular education. In response to the Holy Roman emperor Joseph of Austria's Edict of Toleration of 1781, Naphtali Wessely welcomed the efforts and issued an urgent call for the reform of Jewish education as a prelude to full emancipation. Purely secular subjects—mathematics, German, and world history and literature—were to take precedence over the traditional Jewish studies. The study of the Bible, since it was generally acknowledged to be a fundamental part of Western culture, was to be emphasized at the expense of the more traditional learning in the Talmud. Following this model, modern Jewish schools were established by Jewish intellectuals and businessmen in several German cities, among them Frankfurt and Hamburg. As its educational activities began to bear fruit in the wide dissemination of secular culture, the Berlin Haskala abandoned the use of Hebrew for German and gradually disintegrated. Unlike Mendelssohn himself, the immediate descendants of his circle and his own children were unable to strike a balance between Jewish and secular culture; their Western education undermined their religious faith and they perceived their identity as Europeans rather than as Jews.

One of Mendelssohn's disciples, David Friedlaender, offered to convert to Christianity without accepting Christian dogma or Christian rites; he felt that both Judaism and Christianity shared the same religious truth but that there was no relation at all between Judaism's ceremonial law and that truth. The offer was refused unless Friedlaender would acknowledge the superiority of Christianity and make an unconditional commitment to it, which he was not prepared to do. Unlike Friedlaender, many others who began by following Mendelssohn chose to leave the Jewish faith as the only way to win full acceptance in the European community of which they felt themselves a part.

In eastern Europe. The Haskala, thus, was quickly played out in central Europe; as an idea its further career was to continue in eastern Europe, particularly in the Russian Empire, where it flourished in the middle third of the 19th century until, as a result of the pogroms of 1881, Jews lost faith in the goodwill of Russians to accept "enlightened" Jews. It was a tenet of the Russian Haskala that the tsar was a benevolent leader who would bestow emancipation upon his Jewish subjects as soon as they proved themselves worthy of it; and that it was the task of the Jews, then, to transform themselves into model citizens, enlightened, unsuperstitious, devoted to secular learning and productive occupations. Following the example of the Berlin Haskala, a Russian Hebrew-language writer, Isaac Baer Levinsohn, published a pamphlet, *Teʿuda be-Yisrael* ("Testimony in Israel") citing the benefits of secular education. At the same time, such writers as Joseph Perl and Isaac Erter, though Orthodox Jews themselves, in virulent satire attacked the superstitious folk customs of the masses and opened the way to the anticlericalism which was to become characteristic of the Russian Haskala. In the 1840s and 1850s the emphasis shifted from satire and attack on the cultural parochialism of the Pale of Settlement (the regions to which the Jews were restricted) to romanticization of life outside the Pale, including periods of the Jewish past. Thus, Hebrew poets and novelists, such as Michal Levensohn and Abraham Mapu, arose on Russian soil to contribute their talents to the creation of a modern Hebrew literature. With the climate of government reforms in the 1860s, the Russian Haskala entered a "positivist" phase, calling for practical social and economic reforms. Hebrew-language journals were established and the Hebrew essay and didactic poetry, calling for religious and cultural reforms, came into their own, particularly at the hands of such stylists as the poet Y.L. Gordon and the essayist Moses Leib Lilienblum. Abandoning the original Hebrew and German orientation of the Russian Haskala, a number of Jewish intellectuals, the most prominent of whom were Yoachim Tarnopol, Osip Rabinovich, and Lev Levanda, became Russifiers, founding Russian-language Jewish weeklies devoted to "patriotism, emancipation, modernism." Like their contemporary fellow Jews in western Europe, they declared themselves to be Russians by nationality and Jews by religious belief alone. In 1863 a group of wealthy Jews in St. Petersburg and Odessa created the Society for the Promotion of Culture among the Jews of Russia for the purpose of educating Jewry into "readiness for citizenship." The goal of all segments of the Russian Haskala in the 1860s and 1870s was to turn Jews into good Russians and to make their Jewishness a matter of personal idiosyncrasy alone. The period of reaction that set in with the pogroms (massacres) of 1881 was to prove how deluded the hopes of the Haskala had been.

The modern Hebrew Renaissance

RELIGIOUS REFORM MOVEMENTS

One element of Westernization that the Haskala had championed was the reform of religion. It began in western Europe during the Napoleonic period (1800–15) when certain aspects of Jewish belief and observance were seen as incompatible with the new position of the Jew in Western society. Napoleon convoked a Sanhedrin (Jewish legislative council) in 1807 to create a new, modern definition of Judaism in its renunciation of Jewish nationhood and national aspirations, its protestations that rabbinic authority was purely spiritual, and its recognition of the priority of civil over religious authorities even in the matters of intermarriage. In areas other than France, the rationale for reform, at least in its early years, was more aesthetic than doctrinal. The external aspects of worship—

i.e., the form of the service—appeared unacceptable to the newly Westernized members of the Jewish bourgeoisie in both Germany and the United States, whose standards of cultural acceptability had been shaped by the surrounding society, and who desired above all to resemble their Gentile peers. Thus, the short-lived Reform temple established in Seesen, by the pioneer German reformer Israel Jacobson, in 1810 enshrined order and dignity of a Protestant type in the service and introduced an organ, sermon, and prayers in German, in place of Hebrew, to create an uplifting spiritual experience. The more radical temple in Hamburg (established 1818) adopted all of Jacobson's reforms and published its own much-abridged prayer book, which deleted almost all the references to the long-awaited restoration of Zion. Reformers in Charleston, South Carolina, introduced similar changes in the synagogue ritual in 1824, for they sought a non-national Judaism similar in form to Protestantism and adapted to the surrounding culture. It was apparent to the reformers that in Western society Judaism would have to divest itself of its alien customs and conform to the cultural and intellectual standards of the new "age of reason."

German Reform Judaism

German Reform in the 1840s became institutionalized, a matter of organized, formal belief and practice, and, at a series of synods held at Brunswick (1844), Frankfurt (1845), and Breslau (1846), it created the first theological rationalization for changes introduced in the previous generation. Judaism, it was declared, had always been a developmental religion that conformed to the demands of the times, and, since the Jews were not now a nation, they were no longer bound by their entire religious–political code of law, but only by the dictates of moral law. Thus, those rituals which stood in the way of full Jewish participation in German social and political life were no longer considered valid expressions of Jewish religious truth. The use of Hebrew in religious services was limited; practices such as the dietary laws and circumcision and all national messianic hopes were discarded upon the altar of the "spirit of the times." Messianism in Reform Judaism was transmuted into active concern for social welfare in the present, and the Jewish role in history became Diaspora-centred, a mission to the Gentiles.

Reform Judaism in the United States

Although Reform was initiated in Europe, it did not enjoy a successful career there, for many central European governments that regulated the existence of religious communities would not recognize more than one form of Judaism in any one locale. Reform did not achieve its greatest success until it was imported into the United States along with the massive German-Jewish immigration of the 1840s and coalesced with earlier American trends toward reform. By 1880 almost all of the 200 synagogues in the United States had become Reform, amalgamating in the Union of American Hebrew Congregations (formed 1873). In 1885 the Reform philosophy was given its most comprehensive formulation in the so-called Pittsburgh Platform, drawn up by a conference of Reform rabbis. This manifesto declared that Judaism was an evolutionary faith, and no longer a national faith, and that it was now to be de-orientalized. While the preservation of historical identity was considered to be beneficial, the maintenance of continuity of tradition was not; the Talmud was to be considered merely as religious literature, and not as legislation. The rationalist principles of the Pittsburgh Platform remained the official philosophy of the American Reform movement until a later generation, seeking to meet different emotional and intellectual needs, reintroduced the concept of Jewish peoplehood into the Columbus Platform of 1937, which also reemphasized Hebrew and traditional liturgy and practices. Classical (19th-century) Reform was very much a late child of the Enlightenment, and by mid-20th century its Enlightenment philosophy appeared antiquated to many Jewish thinkers.

Conservative Judaism

If Reform was a child of Enlightenment rationalism, Conservative Judaism was a child of historical romanticism. It began in 1845 when Zacharias Frankel and a group of followers seceded from a second Reform synod at Frankfurt over the issue of the limitation of the use of Hebrew to a small core of prayers. For Frankel, Hebrew represented the spirit of Judaism and the Jewish people, and Judaism itself was not merely a theology of ethics but the historical expression of the Jewish experience; this definition he called "positive-historical Judaism." Although Conservative Judaism conceived of Judaism as a developmental religion, it charted its course through close study of the tradition and the will of the people, and thus came to largely traditional conclusions about religious observance.

ORTHODOX DEVELOPMENTS

In western and central Europe. The bulk of the official Jewish establishment in western and central Europe, though affected by the efforts at religious reform, remained Orthodox (a term first used by Reform leaders to designate their traditionalist opponents). Under the leadership of Samson Raphael Hirsch, a more modern and militant form of Neo-Orthodoxy arose, based in Frankfurt am Main, which asserted its right to break with any Jewish community that contained Reform elements and to form an independent community. The thinking of this group was profoundly influential, for it indicated the possibility of living a ritually and religiously full life while being totally integrated into Western society. It accomplished this by positing a theoretical division between religion and culture; the Jews were to remain Orthodox in religion (although deferring their messianic aspirations to the unforeseeable future) while becoming Western in manners and culture. This form of Orthodoxy, which became the intellectual model for Western Orthodoxy, continued into the late 20th century in the United States in a variety of religious and academic institutions (such as the Yeshiva University and the bulk of English-speaking Orthodox synagogues), coexisting in substantial tension with a number of Orthodox groups, most notably the Lubavitcher and Satmar Ḥasidim (for Ḥasidism, see below), and some Talmudic academies that saw the Western world itself as the enemy and chose to recreate the ghetto.

In eastern Europe. By the mid-18th century Orthodoxy in eastern Europe, having been convulsed by frantic messianism and stifled by the sterility of purely legalistic scholarship, was ripe for revival. The experience of Shabbetaianism (the first messianic movement to excite virtually all of world Jewry) had revealed in the mid-17th century the pervasiveness of Jewish exhaustion with the Exile and fervent longing for messianic redemption, while the nihilistic sect of Frankists (the followers of Jacob Frank, see above *Rabbinic Judaism*) in the 18th century had transmuted that messianism into a this-worldly hysteria. Talmudic piety and study, sunk in excessive *pilpul* (acute logical distinctions that often became mere hairsplitting), was refreshed by the new critical methods of Elijah ben Solomon, the *gaon* of Vilna. Although essentially a legal rigorist, he was open to Western scientific learning insofar as it helped him to elucidate Talmudic texts. Orthodox religious expression also was raised to a new level with the development of Ḥasidism (Pietism) by Israel Baal Shem Tov in the mid-18th century. Although Ḥasidism contained elements of social protest, being at least in part a movement of the poor against the wealthy communal leadership and of the unlearned against the learned—though many of its leaders, among them Rabbi Baer, the *maggid* ("preacher") of Mezhirich and Rabbi Levi Isaac of Berdichev, were well-versed in Talmudic learning—it was essentially a non-messianic outcry in the name of religious emotion, emphasizing prayer and personal religious devotion here and now. Contemporary scholarship is investigating the linkage between Ḥasidism and eastern European Christian pietistic movements. The major innovation that Ḥasidism introduced into Jewish religious life was the charismatic leader, the *rebbe* who served as teacher, confessor, wonder-worker, God's vicar on earth, and occasionally as atoning sacrifice. Although the earliest *rebbes* were democratically chosen, the position of leadership passed to their descendants on the presumption that they had inherited their fathers' charisma and thus created spiritual dynasties. Ḥasidism spread throughout eastern Europe and enjoyed its greatest success in Poland.

Ḥasidism

Ḥasidism was notably unsuccessful, however, in Lithuania, where the traditional rabbinic class, under the leadership of Elijah, the Vilna *gaon*, was able to stave off its influence by issuing a ban of excommunication (*ḥerem*, "anathema") against the new movement. The tactic (a complete boycott and cutting off of communication) was widely embraced by non-Ḥasidic rabbis, who earned for themselves among the Ḥasidim the title of Mitnaggedim (Opponents), but it proved largely ineffective in areas where the rabbis had lost the respect of the masses, and it called forth a round of counter-excommunications by the Ḥasidic *rebbes*. With the passage of time, Ḥasidim and Mitnaggedim abandoned their conflict and came to see each other as allies against the threat to all Orthodox Jewish religion of Haskala and secularization. The impact of Ḥasidism on eastern European Jewry cannot be overemphasized; even in Lithuania, where it did not take firm hold, it stimulated the growth of a home-grown pietism in the Musar (ethicist) movement of the mid-19th century, and it renewed the Talmudic energies of its opponents.

DEVELOPMENTS IN SCHOLARSHIP

As Jews moved into Western society in central Europe, there arose a group of young Jewish intellectuals who devoted themselves to Jewish scholarship of a far different type from traditional Talmudic learning or medieval philosophy. In 1819 Leopold Zunz and Moses Moser founded the Society for Jewish Culture and Learning for the study of Jewish history and literature. Although the original group quickly dissolved, Zunz became the unofficial leader of a generation of scholars dedicated to the Wissenschaft des Judentums (Science of Judaism). Under its carefully objective and scholarly facade, the Wissenschaft movement embraced a variety of nonacademic motives and goals. All of its members sought to prove that the Jewish past was intellectually respectable and worthy of study, and hence that the Jews deserved an equal place within European societies. Jewish scholarship was enlisted as a weapon in the battles for change. Thus, Isaac M. Jost wrote a general history of the Jews to promote Reform; Zunz's *Gottesdienstliche Vorträge der Juden, historisch entwickelt* (1832; "The Worship Sermons of the Jews, Historically Developed") served to legitimize the modern innovation of the sermon in the vernacular; and Abraham Geiger, the outstanding leader of German Reform in the 1840s and 1850s, interpreted the Pharisees as the forerunners of the reformers of his own day. In their work these intellectuals presented archetypes of what modern Jews should become. To support their claims of academic respectability, the Wissenschaft figures highlighted those aspects of the Jewish past that were closely integrated with general fields of study. In particular, Moritz Steinschneider, who owes his fame to towering achievements in bibliography, was concerned above all with the contribution of Jews to science, medicine, and mathematics. Nineteenth-century Jewish scholarship set out to praise Judaism as one of the cofounders of the Western tradition, and thus to argue that whenever the Jews were not excluded from European society they have produced great culture, and that they would repeat such accomplishments under conditions of social and political equality.

The Science of Judaism

The Wissenschaft movement stimulated the critical study of the Jewish past, and great works of synthesis, written from a variety of perspectives, began to appear: Heinrich Graetz's multivolumed *Geschichte der Juden von den ältesten Zeiten bis auf die Gegenwart* (1853–75; *History of the Jews*), written from a romantic-national point of view; Isaac Halevy's *Dorot ha-rishonim* (1897–1932; "The First Generations") and Ze'ev Jawitz's *Toldot Yisrael* (1894; "History of Israel"), from an orthodox standpoint; and Simon Dubnow's *Weltgeschichte des jüdischen Volkes* (1925–29; "World History of the Jewish People"), reflecting his belief in secular, nationalistic communal autonomy. Since the 1920s this tradition of great synthesis has been carried on in the United States by Salo W. Baron, who by the early 1970s had produced 14 volumes of his *Social and Religious History of the Jews* (1952), and in Israel by Benzion Dinur, whose chief work was *Yisrael ba-gola* (3rd ed., 5 vol., 1961–66; "Israel in the Exile"). Many other first-rank scholars in Europe, Israel, and the United States have made notable contributions to the study of Jewish history, rabbinics, and mysticism. This great emphasis on historical research and knowledge from a wide variety of perspectives tended to propose veneration for the Jewish past as a substitute for waning religious faith.

JEWISH–CHRISTIAN RELATIONS

Jewish–Christian relations in the 19th century, strained at best, often erupted into open conflict. Established Christianity, and Roman Catholicism in particular, staunch upholders of the old order, identified the Jews as the major beneficiaries of the French Revolution and as the bearers of a liberal, secular, anticlerical, and often revolutionary threat. Clerical anti-Semitism was thus in France allied with the anti-Semitism of the traditional right, and these movements contended with those who affirmed the results of the French Revolution in the great convulsion of the "Dreyfus affair" in the last years of the 19th century. In Russia the conflict of the Jews and the Orthodox Church released the most open and virulent manifestation of religious anti-Semitism. To the church, the Jews were the enemy seeking to undermine Russian Orthodoxy and the tsar, the very foundations of Russian tradition. The church and the tsarist authorities went so far as to condone, and even encourage, the violent pogroms that were perpetrated against the Jews in 1881–82 and again in 1905. Russian Orthodoxy was active as well in spreading the so-called blood libel, a superstitious belief in Jewish ritual murder which had reemerged even in the 19th century, in Damascus in 1840 (in which instance the French Consul in Syria initiated the accusation) and in Hungary in the Tiszaeszlár affair in 1882. In both cases torture was used to obtain false confessions but the accused were ultimately cleared. The most infamous recurrence of the blood libel in modern times, however, was the Beilis case of 1911–13, in which the tsarist government, with church complicity, sought unsuccessfully to convict a Jewish bookkeeper in Odessa of ritual murder. From Russian Orthodox circles, too, arose the *Protocols of the Learned Elders of Zion*, a fraudulent documentation of an alleged international Jewish conspiracy to conquer the world by subverting the social order through Liberalism, Freemasonry, and other modern movements; the concoction appeared around the turn of the century and enjoyed a phenomenal success in anti-Semitic propaganda. In spite of the fact that much of modern anti-Semitism was not Christian but racial, pagan, and often left-wing, Jews have attributed even secular anti-Semitism to older Christian teachings, which they assert persisted as a case of time lag. In the 20th century Jews and Christians have moved toward mutual understanding. In the early decades of the century, some liberal Christian voices were raised against anti-Semitism; in the United States the National Conference of Christians and Jews was founded (1928) as a response to anti-Semitism propagated in Henry Ford's *Dearborn Independent*. Elements of the Church spoke out during the 1930s against the Nazi persecution of the Jews, but the majority of Christian religious figures in Europe remained silent, even during the Holocaust (near extermination of European Jews). In response to the Holocaust, however, the World Council of Churches denounced anti-Semitism in 1946, and in 1965 the Roman Catholic Church's Schema on the Jews and other non-Christian religions, adopted by the Second Vatican Council, revised its traditional attitude toward the Jews as the killers of Christ. A growing sense of ecumenism (of fellowship and common concerns) has been shared by Jews and Christians alike. Although there remain many difficulties related to the question of the place that Zionism and the State of Israel hold within Judaism, the older forms of official church anti-Semitism have radically lessened.

Christianity and Russian anti-Semitism

Christianity and the Nazi persecution

ZIONISM

The most striking of the new phenomena in Jewish life is Zionism, which, insofar as it has focussed on the return to Zion (the poetic term for the Holy Land), is a re-echo of older religious themes. Insofar as it has stressed the national concentration of the Jews in a secular state, however, it is yet another example of the secularization of Jewish life and of Jewish messianism. In its secular aspects Zionism attempted to complete the emancipation of the

Jews by transforming them into a nation like all other nations. Although it drew upon the general currents of 19th-century European nationalism, its major impetus came from the revival of a virulent form of racist anti-Semitism in the last decades of the 19th century, as noted above. Zionism reacted to anti-Semitic contentions that the Jews were aliens in European society and could never hope to be integrated into it in any numbers, and transformed this charge into a basic premise of a program of national regeneration and resettlement. Zionism has come to occupy roughly the same place in Jewish life as the "social gospel"—according to which the Kingdom of God is to be achieved in economic and social life—for Christians; the involvement in Israel as the new centre of Jewish energies, creativity, and renewal serves as the secular religion of many Diaspora Jews.

AMERICAN JUDAISM

The story of Judaism in the United States is the story of several fresh beginnings. In the colonial period the style of the tiny American Jewish community was shaped by the earliest Sefardic immigrants; the community was officially Orthodox but, unlike European Jewish communities, was voluntaristic, and by the early 19th century there was a significant drift of the younger generation from Judaism. By the mid-19th century a new wave of central European immigrants revived the declining American Jewish community and remade it to serve its own needs. Primarily petty shopkeepers and traders, the new immigrants migrated westward, founding new Jewish centres which were almost entirely controlled by laymen. The exigencies of life on the frontier within an open society created a predisposition for religious reform, and it is significant that the greatest American Reform Jewish leader of the 19th century, Isaac Mayer Wise, was based in Cincinnati, Ohio. Wise sought to unite all of American Jewry in the new nontraditional institutions that he founded, Hebrew Union College (1875), the Union of American Hebrew Congregations (1873), and the Central Conference of American Rabbis (1889); but his ever more radical reforming spirit ultimately drove the traditionalist elements within the American Jewish community into opposition. The head of the traditionalist circles had been Isaac Leeser, a native of Germany, who had attempted to create an indigenous American community on the lines of a modernized traditionalism. Conservative forces, after his death, became disorganized, but in reaction to Reform they defined themselves by their attachment to the sabbath, the dietary laws, and especially to Hebrew as the language of prayer. Under the leadership of Sabato Morais, an Orthodox Jew of Italian birth, Conservative circles in 1886 founded a rabbinic seminary of their own, the Jewish Theological Seminary of America.

The central and eastern European immigrant styles of Judaism

The eastern European immigrants who moved in large numbers to the American shores in the years from 1881 to 1914 were profoundly different in culture and manners from the older elements of the American Jewish community, and it is they and their descendants who made American Judaism as it is today. The bridge between the existing Jewish community led by German Jews of Reform persuasion and the new immigrant masses was the traditionalist element among the older settlers. Cyrus Adler, traditionalist himself, cooperated with a German Reform circle of Jacob Schiff in reorganizing the Jewish Theological Seminary (1902) and other institutions for the purpose of Americanizing the eastern European immigrants. Enough eastern European rabbis and scholars had immigrated, however, to create their own synagogues, which reproduced the customs of the old world. Whereas in 1880 almost all of the 200 Jewish congregations in the United States were Reform, by 1890 there were 533 synagogues, and most of the new ones founded by immigrant groups were Orthodox. The Union of Orthodox Jewish Congregations, which was established in 1898 by elements associated with the Jewish Theological Seminary, was soon taken over by Yiddish-speaking recent immigrants for whom the seminary was much too liberal. In 1902 immigrant rabbis also formed their own body, the Union of Orthodox Rabbis of the United States and Canada (the Agudath ha-Rabbanim), which fostered the creation of yeshivas (rabbinic academies) of the old type. In 1915 two small yeshivas, Etz Chaim and Rabbi Isaac Elhanan Theological Seminary, added Yeshiva College of secular studies in 1928, and became Yeshiva University in 1945. The eastern European Orthodox elements concentrated primarily on Jewish education and it was they who introduced the movement for Jewish day schools, analogous to Christian parochial schools. Gradually an American version of Orthodoxy developed on the Neo-Orthodox model of Samson Raphael Hirsch, which combined institutional separatism and cooperation with other Jewish groups in umbrella organizations.

American Orthodoxy

The immigrants and their children had three desires; to upgrade themselves socially by joining older congregations or forming their own in an Americanized image; to affirm an unideological commitment to Jewish life; and to maintain their ties to the overseas Jewish communities of their origin. With their strong sense of Jewish peoplehood, they introduced Zionism into American Jewish life, and accepted the basic ideas of Mordecai Kaplan's Reconstructionism, which was committed to Zionism. A small group of anti-Zionists remained a significant force in the 1930s and 1940s, but their central organization, the American Council for Judaism, represented the descendants of earlier German-Jewish immigrants. The later immigrants took over all the earlier institutions of the Jewish community and imbued them with their own spirit.

The religious continuum

American Jewish religious life is a continuum—from the most traditional Orthodoxy to the most radical Reconstructionism. In theory, all of the Orthodox groups agree on the revealed nature of all of Jewish law; for the Reform groups, the moral doctrine of Judaism is divine and its ritual law is man made; the Conservatives see Judaism as the working out in both areas of a divine revelation that is incarnate in a slowly changing human history; and the Reconstructionists (who include both Conservative and Reform Jews) view Judaism as the evolving civilization created by the Jewish people in the light of its highest conscience. What really marks the various bodies in the mind of the Jewish community is their difference in ritual practices, but the ritual variations shade from one group into the other. The role of the rabbi is substantially the same in all three groups; he is no longer a Talmudic scholar but a preacher, pastor, and administrator, a cross between a parish priest and the leader of an ethnic group. Although there was some cooperation among the three major Jewish denominations—Orthodoxy, Reform, and Conservatism—the real effort of organized Jewish religion in America in the late 20th century revolved around the individual synagogue and the denomination to which it belonged. As religious identification became increasingly respectable in American life, the Jews followed the American norm, affiliating in greater numbers with synagogues, though often for ethnic or social, rather than religious reasons.

JUDAISM IN OTHER LANDS

Modernity came first to the Jewish people in Europe and it was, therefore, within the European context that representatives of important non-Ashkenazi communities such as the proto-Zionist Sefardi Judah ben Solomon Ḥai Alkalai of Sarajevo, the Luzzatto family in Italy, and Elijah ben Abraham Benamozegh in France, participated in variations of Jewish modernity. In England and France, more than in Germany or Russia, Wissenschaft des Judentums (see above *Developments in scholarship*), with its enlightenment ideology, was the central focus of Jewish experience; there the "republic of scholarship" became the synagogue of the Jewish intelligentsia. In neither country did Reform Judaism gain a major foothold, for the Orthodox establishment, which remained the official synagogue, liberalized its synagogue practice while retaining its essentially conservative outlook. In Anglo-Jewish life in the last decades of the 19th century the two most pronounced modernist tendencies were Solomon Schechter's moderate romantic traditionalism and the "renewed Karaism" of Claude Joseph Goldsmid Montefiore, whose version of religious reform was "back to the Bible."

Outside of Europe, in such places as South America and Canada, Jewish modernity appeared late, for European Jewry arrived in those places even later than in the United States, attaining a significant number only in the 20th century. These communities have been dependent on immigrant scholars and intellectuals for serious Jewish thought. Jews in the Arab lands, in North Africa and the Middle East, living within traditional societies, entered modernity even later than those on the peripheries of Europe. Many of them received their first introduction to the Western world in widespread schools set up by the Alliance Israélite Universelle (a Jewish defense organization centred in Paris), which combined Jewish education with the language and values of French civilization. Yet most of these communities remained traditionalist almost to the moment when they were expelled or felt compelled to relocate, since 1948, when the state of Israel was created. The ferment of modernity in all its forms is now being felt in their ranks. In Israel, which has received a large segment of Sefardic Jewry, the attention of these communities has turned to attaining equality with the more advanced Ashkenazim rather than developing some forms of modern Jewish thought.

Other groups that may be described as regional or ethnic include the Bene-Israel, descendants of Jewish settlers in the Bombay region of India, whose deviation in some Halakhic matters from the present Orthodox consensus has raised problems for those among them who have migrated to Israel; the Falashas, the Jewish community of Ethiopia whose development has been almost entirely outside the mainstream described in this article; and the Black Jews of the United States, whose place in, and relation to, the rest of the community remains unclear.

CONTEMPORARY JUDAISM

As a result of the Holocaust, Judaism has become a non-European religion; its three major centres, which together include more than three-fourths of world Jewry, are Israel, the Slavic region of the former Soviet Union, and the United States. Although Jews constitute only a small fraction of the population of the United States, Judaism occupies a role far surpassing its numerical importance and is regarded with Roman Catholicism and Protestantism as one of the major American faiths. Similarly, in the international realm of Western religion, Judaism has been welcomed as a partner able to deal with other major religions as an equal on such issues as anti-Semitism, human rights, and world peace.

The rights and needs of the world Jewish community, including Israel, have triggered deep conflicts with which Judaism has been involved with the Arab and Communist worlds. Friction between Israel and the Arab states has created tension with Islām, while the political stand of Israel and the treatment of the Jewish minority within the erstwhile Soviet Union led to open clashes with the Communist leadership. Some of the diatribes and charges that have issued from the Arabs and Communists in this struggle have at times re-evoked older forms of anti-Semitism. In the long-range view, the problems of Judaism and Islām seem more soluble than those of Judaism and such secular ideologies as Communism, for the major religions of the world are increasingly seeking accommodation with each other, as all are confronted by hostile secularist ideologies which have retained their conversionary élan.

Secularization and Jewish identity

Within its own community Jewry is faced with the increasing secularization of Jewish identity in its three major centres, each in its own way. In the United States the open society and the melting pot ideologies of past generations have fostered among many Jews a sense of Jewish identity increasingly devoid of concrete religious, national, or historical content; in the former Soviet Union government policy from the 1930s had banned the teaching of Judaism and Jewish culture to the young and had severely discouraged any manifestation of Jewish identity as a sign of the disloyalty of "rootless cosmopolitans" to the Russian state; and in Israel a secular nationalism has taken root, raising questions as to the role that Judaism plays in the identity of the average Israeli. Nonetheless, underneath the external secularization there are signs of a persisting deep Jewish religious fervour, in which the sense of history, community, and personal authenticity figure as the intertwined strands of Jewish religious life, especially as it has been affected by the State of Israel. Some of the rituals of the Jewish tradition, especially the rites of passage at the crucial stages of individual existence, are almost universally observed; in the United States, for example, more than 80 percent of Jewish children receive some formal religious training. Among Jewish youth there is, in some circles, a quest for tradition. In the United States, Jewish communes have been established that seek new forms of Jewish expression; in Israel, groups such as Mevaqshe Derekh (Seekers of the Way) have tried to bridge secular Israeli culture and Jewish tradition and to maintain traditional Jewish ethical standards even in wartime; in Russia thousands of young people gather on several occasions of the year to dance and sing and express solidarity in front of the synagogues in St. Petersburg and Moscow. Still, signs of major weaknesses persist. The rate of intermarriage among Jews in the Diaspora increased, while regular synagogue attendance, at the very highest 20 percent in the United States, remained far below church attendance. Despite their lack of traditional piety, there is a general sense among Jews that they remain Jews not because of the force of anti-Semitism but because of the attractiveness of their tradition and their sense of a common history and destiny.

The enduring existence of the Jews and Judaism

If in 1945, the world Jewish community, decimated and horrified by the Holocaust, felt in danger of disappearing, there appeared to be no such despair in the last quarter of the century when there was an expectation that Jewish communal feeling would remain strong, especially, for many or most Jews, in light of the existence of Israel. Judaism enjoyed a heightened dignity in the eyes of the world, not only because of the creation of the State of Israel, but also because of its close relations with other world religions. Although the recurring phenomenon of the alienation of young Jews from their tradition was troubling, it is no more so than in recent past generations. Along with other major religions, Judaism's most disturbing problem yet to be solved was how to deal with secular ideologies and the growth of secularism within its own ranks. Thus, in the last decades of the 20th century, it appeared that Judaism would have to contend with as many problems as other major religions, but that it faced them with no less confidence than these, and with more confidence than it had felt earlier in the century. (Ar.H.)

THE JUDAIC TRADITION

The literature of Judaism

A paradigmatic statement is made in the narrative that begins with Genesis and concludes with Joshua. In the early chapters of Genesis the divine is described as Creator of the natural order, including mankind. In the Eden, Flood, and Tower of Babel stories, man is recognized as rebellious and disobedient. In the patriarchal stories (about Abraham, Isaac, Jacob, and Joseph) a particular family is called out of humankind to restore the thwarted relationship through personal and communal responsiveness. The subsequent history of the community thus formed is recounted so that the divinely sought restoration may be recognized and the nature of the obedient community may be observed: the Egyptian servitude, the going out from Egypt, the revelation of the "teaching," the wandering years, and finally fulfillment through entrance into the "land" (Canaan). The prophetic books (in the Hebrew Bible these include the historical narratives up to the Babylonian Exile—*i.e.,* Joshua, Judges, Samuel, and Kings) continue to deal with the rebellion-obedience tension, interpreting it within the changing historical context and adding new levels of meaning to the fulfillment-redemption motif. (The literature of the Old Testament is treated in the article BIBLICAL LITERATURE.)

Philosophical, mystical, and exegetical additions

It is from this "narrative theology," as it has been recited throughout the centuries, that new formulations of the primal affirmations have been drawn. These have been clothed in a number of vocabularies: philosophical, mystical, ethnic, political, and others. The emphases have been various, the disagreements often profound. No single exposition has exhausted the possibilities of the affirmations or of the relationship between them. Philosophers have expounded them on the highest level of abstraction, using the language of the available philosophic systems. Mystics have enveloped them in the extravagant prose of speculative systems and in simple folktales. Attempts have been made to encompass them in theoretical ethical statements and express them through practical ethical behaviour. Yet, in each instance, the proposed interpretations have had to come to terms with the biblical affirmations and with the particular mode of understanding them required by the spiritual and intellectual demands arising out of the community's experience. The biblical texts, themselves the products of a long period of transmission and embodying more than a single outlook, were subjected to extensive study and interpretation over many centuries and, when required, were translated into other languages. The whole literature continued to provide the basis of further developments, so that any attempt to formulate a statement of the affirmations of Judaism must, however contemporary it seeks to be, give heed to the scope and variety of speculation and formulation in the past.

SOURCES AND SCOPE OF THE TORAH

The concept "Giver of Torah" played a central role in the understanding of God, for it is Torah, or "teaching," that confirms the events recognized by the community as the act of God. In its written form, Torah was considered to be especially present in the first five books of the Bible (the Pentateuch), which therefore came to be called Torah. In addition to this written Torah, or "Law," there were also unwritten laws or customs and interpretations of them, carried down in an oral tradition over many generations, which acquired the status of oral Torah.

The Talmud ("study" or "learning") is the literary culmination of this oral tradition, which, according to the rabbis who created the Talmud, originated at Mt. Sinai as part of the divine revelation vouchsafed to Moses, along with the material recorded in the Pentateuch. In its broadest sense, the Talmud is a set of books consisting of the Mishna ("repeated study"), the Gemara ("completion"), and certain auxiliary materials. The Mishna is a collection of originally oral laws supplementing scriptural laws. The Gemara is a collection of commentaries on and elaborations of the Mishna, which in "the Talmud" is reproduced in juxtaposition to the Gemara. For present-day scholarship, however, Talmud in the precise sense refers only to the materials customarily called Gemara—an Aramaic term prevalent in medieval rabbinic literature that was used by the church censor to replace the term Talmud within the Talmudic discourse in the Basel edition of the Talmud, published 1578–81. This practice continued in all later editions.

Mishna and Gemara

The term Midrash ("exposition" or "investigation"; plural, Midrashim) is also used in two senses. On the one hand, it refers to a mode of biblical interpretation prominent in the Talmudic literature; on the other, it refers to a separate body of commentaries on Scripture using this interpretative mode.

The oral tradition interpreted the written Torah, adapted its precepts to ever-changing political and social circumstances, and supplemented it with new legislation. Thus the oral tradition added a dynamic dimension to the written code, making it a self-regenerating, endless source of guidance, a perpetual process rather than a closed system. The vitality of this tradition is fully demonstrated in the way the ancient laws were adapted after the destruction of the Temple in 70 CE and by the role the Talmud played in the survival of the Jewish people in exile. By the 11th century, Diaspora Jews lived within a Talmudic culture that unified them and that superseded geographical boundaries and language differences. Jewish communities governed themselves according to Talmudic law, and individuals regulated the smallest details of their lives by it.

Central to this vast structure was, of course, the Jewish community's concern to live in accordance with the divine will embodied and expressed in Torah in the widest sense. Scripture, Halakhic and Haggadic Midrash, Mishna, and Gemara were the sources from which the leaders of the communities drew in order to provide both stability and flexibility. The dispersion of the Jews outside Palestine confronted communities and individuals with novel and unexpected situations that had to be dealt with in such a way as to provide continuity while at the same time making it possible to exist with the unprecedented.

Prophecy and religious experience. Torah in the broad sense includes the whole Hebrew Bible, including the prophetic books. In biblical prophecy, God is seen as continuing to be disclosed in the nexus of historical events and as making ethical demands upon the community. According to rabbinic Judaism, this source of Torah—the charismatic person—dried up in the period of Ezra (*i.e.*, about the time of the return from the Babylonian Exile in the 5th century BCE). This opinion may have been a defensive reaction to the luxuriant growth of apocalyptic speculation about the end of the world and the kingdoms of this world, a development that was considered dangerous and unsettling in the period after the Bar Kokhba revolt (132–135 CE). Indeed, there appears to have developed an ongoing suspicion that unrestrained individual experience as the source of Torah was inimical to the welfare of the community. Such an attitude was by no means new. Deuteronomy (13:2–19) had already warned against such "misleaders." The culmination of this attitude is to be found in a Talmudic narrative in which even the *bat qol*, the divine "echo" that announces God's will, is ignored on a particular occasion. Related to this is the reluctance on the part of teachers in the early Christian centuries to point to wonders and miracles in their own time. Far from expressing an ossification of religious experience—the development of the *Siddur* and the Talmudic reports on the devotional life of the rabbis contradict such an interpretation—the attitude seems to be a response to the development of such religious enthusiasm as was exhibited, for example, in the behaviour of the Christian Church in Corinth (I Cor.) and among Gnostic sects and sectarians. Thus, even among the speculative mystics of the Middle Ages, where allegorization of Scripture abounds, the structure of the community and the obligations of the individual are not displaced by the deepening of personal religious life through mystical experience. The decisive instance of this is Joseph Karo (16th century), who was thought to be in touch with a supernal guide, but who was, at the same time, the author of an important codification of Jewish law, the *Shulḥan ʿarukh.*

Anticharismatic trends in post-biblical Judaism

Admittedly, there have been occasions when Torah, even in the wide sense, has been rigidly viewed and applied. In certain historical situations, the dynamic process of rabbinic Judaism has been treated as a static structure. What is of greater significance, however, is the way in which this tendency toward inflexibility has been checked and reversed by the inherent dynamism of the rabbinic tradition.

Modern views of Torah. In modern times—since the end of the 18th century—the traditional position has been challenged both in detail and in principle. The rise of biblical criticism has raised a host of questions about the origins and development of Scripture and thus about the very concept of Torah, in the senses in which it has functioned in Judaism. Naturalistic views of God have required a reinterpretation of Torah in sociological terms as the ideals and sancta (holy things) of the Jewish people. Other and varying positions of many sorts have been and undoubtedly will be forthcoming. What is crucial, however, is the concern of all these positions to retain—with whatever modifications are required—the concept of Torah as one of the central and continuing affirmations of Judaism.

Opposition to the Talmud. Despite the central place of the Talmud in traditional Jewish life and thought, sig-

Anti-Talmudic positions

nificant Jewish groups and individuals have opposed it vigorously. The Karaite sect in Babylonia, beginning in the 8th century, refuted the oral tradition and denounced the Talmud as a rabbinic fabrication. Medieval Jewish mystics declared the Talmud a mere shell covering the concealed meaning of the written Torah, and heretical messianic sects in the 17th and 18th centuries totally rejected it. The decisive blow to Talmudic authority came in the 18th and 19th centuries when the Haskala (the Jewish Enlightenment movement) and its aftermath, Reform Judaism, secularized Jewish life and, in doing so, shattered the Talmudic wall that had surrounded the Jews. Thereafter, modernized Jews usually rejected the Talmud as a medieval anachronism, denouncing it as legalistic, casuistic, devitalized, and unspiritual.

There is also a long-standing anti-Talmudic tradition among Christians. The Talmud was frequently attacked by the church, particularly during the Middle Ages, and accused of falsifying biblical meaning, thus preventing Jews from becoming Christians. The church held that the Talmud contained blasphemous remarks against Jesus and Christianity and that it preached moral and social bias toward non-Jews. On numerous occasions the Talmud was publicly burned, and permanent Talmudic censorship was established.

On the other hand, since the Renaissance there has been a positive response and great interest in rabbinic literature by eminent non-Jewish scholars, writers, and thinkers in the West. As a result, rabbinic ideas, images, and lore, embodied in the Talmud, have permeated Western thought and culture.

Content, style, and form. The Talmud is first and foremost a legal compilation. At the same time it contains materials that encompass virtually the entire scope of subject matter explored in antiquity. Included are topics as diverse as agriculture, architecture, astrology, astronomy, dream interpretation, ethics, fables, folklore, geography, history, legend, magic, mathematics, medicine, metaphysics, natural sciences, proverbs, theology, and theosophy.

This encyclopaedic array is presented in a unique dialectic style that faithfully reflects the spirit of free give-and-take prevalent in the Talmudic academies, where study was focussed upon a Talmudic text. All present participated in an effort to exhaust the meaning and ramifications of the text, debating and arguing together. The mention of a name, situation, or idea often led to the introduction of a story or legend that lightened the mood of a complex argument and carried discussion further.

This text-centred approach profoundly affected the thinking and literary style of the rabbis. Study became synonymous with active interpretation rather than with passive absorption. Thinking was stimulated by textual examination. Even original ideas were expressed in the form of textual interpretations.

Halakha and Haggada; Midrash and Mishna

The subject matter of the oral Torah is classified according to its content into Halakha and Haggada and according to its literary form into Midrash and Mishna. Halakha ("law") deals with the legal, ritual, and doctrinal parts of Scripture, showing how the laws of the written Torah should be applied in life. Haggada ("narrative") expounds on the nonlegal parts of Scripture, illustrating biblical narrative, supplementing its stories, and exploring its ideas. The term Midrash denotes the exegetical method by which the oral tradition interprets and elaborates scriptural text. It refers also to the large collections of Halakhic and Haggadic materials that take the form of a running commentary on the Bible and that were deduced from Scripture by this exegetical method. In short, it also refers to a body of writings. Mishna is the comprehensive compendium that presents the legal content of the oral tradition independently of scriptural text.

Modes of interpretation and thought. Midrash was initially a philogical method of interpreting the literal meaning of biblical texts. In time it developed into a sophisticated interpretive system that reconciled apparent biblical contradictions, established the scriptural basis of new laws, and enriched biblical content with new meaning. Midrashic creativity reached its peak in the schools of Rabbi Ishmael and Akiba, where two different hermeneutic methods were applied. The first was primarily logically oriented, making inferences based upon similarity of content and analogy. The second rested largely upon textual scrutiny, assuming that words and letters that seem superfluous teach something not openly stated in the text.

The Talmud (*i.e.*, the Gemara) quotes abundantly from all Midrashic collections and concurrently uses all rules employed by both the logical and textual schools; moreover, the Talmud's interpretation of Mishna is itself an adaptation of the Midrashic method. The Talmud treats the Mishna in the same way that Midrash treats Scripture. Contradictions are explained through reinterpretation. New problems are solved logically by analogy or textually by careful scrutiny of verbal superfluity.

Midrashic-Talmudic thinking

The strong involvement with hermeneutic exegesis—interpretation according to systematic rules or principles—helped develop the analytic skill and inductive reasoning of the rabbis but inhibited the growth of independent abstract thinking. Bound to a text, they never attempted to formulate their ideas into the type of unified system characteristic of Greek philosophy. Unlike the philosophers, they approached the abstract only by way of the concrete. Events or texts stimulated them to form concepts. These concepts were not defined but, once brought to life, continued to grow and change meaning with usage and in different contexts. This process of conceptual development has been described by some as "organic thinking." Others use this term in a wider sense, pointing out that, although rabbinic concepts are not hierarchically ordered, they have a pattern-like organic coherence. The meaning of each concept is dependent upon the total pattern of concepts, for the idea content of each grows richer as it interweaves with the others.

EARLY COMPILATIONS

Ezra the scribe who, according to the Book of Ezra, reestablished and reformed the Jewish religion in the 5th century BCE, began the "search in the Law . . . to teach in Israel statutes and ordinances."

His work was continued by *soferim* (scribes), who preserved, taught, and interpreted the Bible. They linked the oral tradition to Scripture, transmitting it as a running commentary on the Bible. For almost 300 years they applied the Torah to changing circumstances, making it a living law. They also introduced numerous laws that were designated "words of the *soferim*" by Talmudic sources. By the end of this period, rabbinic Judaism—the religious system constructed by the scribes and rabbis—was strong enough to withstand pressure from without and mature enough to permit internal diversity of opinion.

At the beginning of the 2nd century BCE, a judicial body headed by the *zugot*—pairs of scholars—assumed Halakhic authority. There were five pairs in all, between *c.* 150 and 30 BCE. The first of the *zugot* also introduced the Mishnaic style of transmitting the oral tradition.

The *tannaim*

The making of the Mishna: 2nd–3rd centuries. Hillel and Shammai, the last of the *zugot*, ushered in the period of the *tannaim*—"teachers" of the Mishna—at the end of the 1st century BCE. This era, distinguished by a continuous attempt to consolidate the fragmentary Midrashic and Mishnaic material, culminated in the compilation of the Mishna at the beginning of the 3rd century CE. The work was carried out in the academies of Hillel and Shammai and in others founded later. Most scholars believe that Halakhic collections existed prior to the fall of Jerusalem, in 70 CE. Other compilations were made at Yavne, a Palestinian town near the Mediterranean, as part of the effort to revitalize Judaism after the disaster of 70 CE. By the beginning of the 2nd century there were many such collections. Tradition has it that Rabbi Akiba organized much of this material into separate collections of Midrash, Mishna, and Haggada and introduced the formal divisions in tannaitic literature. His students and other scholars organized new compilations that were studied in the different academies.

After the rebellion of the Jews against Roman rule led by Simeon bar Kokhba in 132–135, when the Sanhedrin (the Jewish supreme court and highest academy) was revived, the Mishnaic compilation adopted by the Sanhedrin president became the official Mishna. The Sanhedrin reached its highest stature under the leadership of Judah ha-Nasi (Judah the Prince, or President); he was also called Rabbi, as the preeminent teacher.

It seems certain that the official Mishna studied during his presidency was the Mishna we know and that he was its editor. Judah aimed to include the entire content of the oral tradition. He drew heavily from the collections of Akiba's pupils but also incorporated material from other compilations, including early ones. Nevertheless, the accumulation was such that selection was necessary. Thus almost no Midrash or Haggada was included. Colleagues and pupils of Judah not only made minor additions to the Mishna but tried to preserve the excluded material, the Baraitot ("Exclusions"), in separate collections. One of these was the Tosefta ("Addition"). Midrashic material was gathered in separate compilations, and later revisions of some of these are still extant. The language of all of the tannaitic literature is the new Hebrew developed during the period of the Second Temple (*c.* 6th century BCE–1st century CE).

The *amoraim*

The making of the Talmuds: 3rd–6th century. The expounders of the Mishna were the *amoraim* ("interpreter"), and the two Talmuds—the Palestinian (or Jerusalem) and the Babylonian—consist of their explanations, discussions, and decisions. Both take the form of a running commentary on the Mishna.

The foundations for these two monumental works were begun by three disciples of Judah ha-Nasi: Joḥanan bar Nappaḥa, Rav (Abba Arika), and Samuel bar Abba, in their academies at Tiberias, in Palestine, and at Sura and Nehardea in Babylonia, respectively. Centres of learning where the Mishna was expounded existed also at Sepphoris, Caesarea, and Lydda in Palestine. In time new academies were established in Babylonia, the best known being those at Pumbedita, Mahoza and Naresh, founded by Judah bar Ezekiel, Rava, and Rav Pappa, respectively. The enrollment of these centres often numbered in the thousands, and students spent many years there. Those who no longer lived on the academy grounds returned twice annually for the *kalla,* a month of study in the spring and fall.

Academies differed in their methods of study. Pumbedita, for example, stressed casuistry, while Sura emphasized breadth of knowledge. Students often moved from one academy to another and even from Palestine to Babylonia or from Babylonia to Palestine. This kept open the channels of communication between the various academies and resulted in the inclusion of much Babylonian material in the Palestinian Talmud, and vice versa.

Despite the overwhelming similarity of the two Talmuds, however, they do differ in some ways. The Palestinian Talmud is written in the Western Aramaic dialect, the Babylonian in the Eastern. The former is invariably shorter, and, not having been subject to final redaction, its discussions are often incomplete. Its explanations tend to remain closer to the literal meaning of the Mishna, preferring textual emendation to casuistic interpretation. Finally, some of the legal concepts in the Babylonian Talmud reflect the influence of Persian law, for Babylonia was under Persian rule at the time.

The main endeavour of the *amoraim* was to thoroughly explain and exhaust the meaning of the Mishna and the Baraitot. Apparent contradictions were reconciled by such means as explaining that conflicting statements referred to different situations or by asserting that they stemmed from the Mishnayot (Mishnas) of different *tannaim.* The same techniques were used when amoraic statements contradicted the Mishna. These discussions took place for hundreds of years, and their content was passed on from generation to generation, until the compilation of the Talmud.

The compilation of the Palestinian and Babylonian Talmuds

The portion of the Palestinian Talmud dealing with the three Bavot ("gates")—*i.e.,* the first three tractates of the fourth order of the Mishna (for orders and tractates, see *Talmudic and Midrashic literature,* below)—was compiled in Caesarea in the middle of the 4th century and is distinguished from the rest by its brevity and terminology. The remainder was completed in Tiberias some 50 years later. It seems likely that its compilation was a rescue operation designed to preserve as much of the Halakhic material collected in Palestinian academies as possible, for by that time the deterioration of the political situation had forced most Palestinian scholars to emigrate to Babylonia.

The Babylonian Talmud was compiled up to the 6th century. Some scholars suggest that the organization of the Talmud began early and that successive generations of *amoraim* added layer upon layer to previously arranged material. Others suggest that at the beginning a stratum called Gemara, consisting only of Halakhic decisions or short comments, was set forth. Still others theorize that no overall arrangement of Talmudic material was made until the end of the 4th century.

The statement in the tractate *Bava metzia* that "Rabina and Rav Ashi were the end of instruction" is most often understood as referring to the final redaction of the Talmud. Since at least two generations of scholars following Rav Ashi (died 427) are mentioned in the Talmud, most scholars suggest that "Rabina" refers to Rabina bar Huna (died 499) and that the redaction was a slow process lasting about 75 years to the end of the 5th century.

According to the tradition of the *geonim*—the heads of the academies at Sura and Pumbedita from the 6th to the 11th centuries—the Babylonian Talmud was completed by the 6th-century *savoraim* ("expositors"). But the extent of their contribution is not precisely known. Some attribute to them only short additions. Others credit them with creating the terminology linking the phases of Talmudic discussions. According to another view, they added comments and often decided between conflicting opinions. The proponents of the so-called Gemara theory noted above ascribe to them the entire dialectic portion of Talmudic discourse.

TALMUDIC AND MIDRASHIC LITERATURE

Mishna. The Mishna is divided into six orders (*sedarim*), each order into tractates (*massekhtot*), and each tractate into chapters (*peraqim*). The six orders are *Zeraʿim, Moʿed, Nashim, Neziqin, Qodashim,* and *Ṭohorot.*

Zeraʿim, Moʿed, and *Nashim*

1. *Zeraʿim* ("Seeds") consists of 11 tractates: *Berakhot, Pea, Demai, Kilayim, Sheviʿit, Terumot, Maʿaserot, Maʿaser sheni, Ḥalla, ʿOrla,* and *Bikkurim.* Except for *Berakhot* ("Blessings"), which treats of daily prayers and grace, this order deals with laws related to agriculture in Palestine. It includes prohibitions against mixtures in plants (hybridization), legislation relating to the sabbatical year (when land lies fallow and debts are remitted), and regulations concerning the portions of harvest given to the poor, the Levites, and the priests.

2. *Moʿed* ("Season" or "Festival") consists of 12 tractates: *Shabbat, ʿEruvin, Pesaḥim, Sheqalim, Yoma, Sukka, Betza, Rosh Hashana, Taʿanit, Megilla, Moʿed qaṭan,* and *Ḥagiga.* This order deals with ceremonies, rituals, observances, and prohibitions relating to special days of the year, including the Sabbath, holidays, and fast days. Since the half-shekel Temple contribution was collected on specified days, tractate *Sheqalim,* regarding this practice, is included.

3. *Nashim* ("Women") consists of seven tractates: *Yevamot, Ketubbot, Nedarim, Nazir, Soṭa, Giṭṭin,* and *Qiddushin.* This order deals with laws concerning betrothal, marriage, sexual and financial relations between husband and wife, adultery, and divorce. Since Nazirite (ascetic) and other vows may affect marital relations, *Nedarim* ("Vows") and *Nazir* ("Nazirite") are included here.

Neziqin, Qodashim, and *Ṭohorot*

4. *Neziqin* ("Damages") consists of 10 tractates, the first three of which were originally considered one (the *Bavot*): *Bava qamma, Bava metzia, Bava batra, Sanhedrin, Makkot, Shevuʿot, ʿEduyyot, ʿAvoda zara, Avot,* and *Horayot.* This order deals with civil and criminal law concerning damages, theft, labour relations, usury, real estate, partnerships, tenant relations, inheritance, court composition, jurisdiction and testimony, erroneous decisions of the Sanhedrin, and capital and other physical punishments. Since idolatry, in the literal sense of worship or veneration of material images, is punishable by death, *ʿAvoda zara* ("Idolatry") is included. *Avot* ("Fathers"), commonly called "Ethics of the Fathers" in English, seems to have

been included to teach a moral way of life that precludes the transgression of law.

5. *Qodashim* ("Sacred Things") consists of 11 tractates: *Zevaḥim, Menaḥot, Ḥullin, Bekhorot, ʿArakhin, Temura, Keretot, Meʿila, Tamid, Middot,* and *Qinnim.* This order incorporates some of the oldest Mishnaic portions. It treats of the Temple and includes regulations concerning sacrifices, offerings, and donations. It also contains a detailed description of the Temple complex.

6. *Ṭohorot* ("Purifications") consists of 12 tractates: *Kelim, Ohalot, Negaʿim, Para, Ṭohorot, Miqwaʾot, Nidda, Makhshirin, Zavim, Ṭevul yom, Yadayim,* and *ʿUqtzin.* This order deals with laws governing the ritual impurity of vessels, dwellings, foods, and persons, and with purification processes.

Tosefta. The Tosefta ("Addition") closely resembles the Mishna in content and order. In its present form it at times supplements the Mishna, at other times comments on it, and often also opposes it. There is no Tosefta on the tractates *Avot, Tamid, Middot,* and *Qinnim.* The Talmud quotes from many other collections of Mishnaiot and Baraitot: some are attributed to *tannaim,* and predate the established Mishna; and others, to *amoraim.* The original material is lost.

Talmud (Gemara). Although the entire Mishna was studied at the Palestinian and Babylonian academies, the Palestinian Talmud (Gemara) covers only the first four orders (except chapters 21–24 of *Shabbat* and chapter 3 of *Makkot*) and the first three chapters of *Nidda* in the sixth order. Most scholars agree that the Palestinian Talmud was never completed to the fifth and sixth orders of the Mishna and that the missing parts of the other orders were lost. A manuscript of chapter 3 of *Makkot* was, in fact, found and was published in 1946.

Incompleteness of the Talmuds

The Babylonian Talmud does not cover orders *Zeraʿim* (except *Berakhot*) and *Ṭohorot* (except *Nidda*) and tractates *Tamid* (except chapters 1,2,4), *Sheqalim, Middot, Qinnim, Avot,* and *ʿEduyyot.* Scholars concur that the Talmud for these parts was never completed, possibly because their content was not relevant in Babylonia.

Midrashim. *Halakhic.* Halakhic Midrashim are exegetic commentaries on the legal content of Exodus, Leviticus, Numbers, and Deuteronomy. The five extant collections are *Mekhilta,* on Exodus; *Mekhilta deRabbi Shimʿon ben Yoḥai,* on Exodus; *Sifra,* on Leviticus; *Sifre,* on Numbers and Deuteronomy; *Sifre zuṭa,* on Numbers. (*Mekhilta* means "measure," a norm or rule; *Sifra,* plural *Sifre,* means "writing" or "book.") Critical analysis reveals that *Mekhilta* and *Sifre* on Numbers differ from the others in terminology and method. Most scholars agree that these two originated in the school of Ishmael and the others in that of Akiba. In their present form they also include later additions. Mention should also be made of *Midrash tannaim* on Deuteronomy, consisting of fragments recovered from the Yemenite anthology *Midrash ha-gadol.*

Haggadic. Haggadic Midrashim originated with the weekly synagogue readings and their accompanying explanations. Although Haggadic collections existed in tannaitic times, extant collections date from the 4th–11th centuries. Midrashic compilations were not authoritatively edited and tend to be coincidental and fragmentary.

Biblical and nonbiblical Haggadot

Most notable among biblical collections is *Midrash rabba* ("Great Midrash"), a composite of commentaries on the Pentateuch and five Megillot (Song of Songs, Ruth, Ecclesiastes, Esther, Lamentations) differing in nature and age. Its oldest portion, the 5th-century *Genesis rabba,* is largely a verse-by-verse commentary, while the 6th-century *Leviticus rabba* consists of homilies and *Lamentations rabba* (end of 6th century) is mainly narrative. The remaining portions of *Midrash rabba* were compiled at later dates.

The *Tanḥuma* (after the late-4th-century Palestinian *amora* Tanḥuma bar Abba), of which two versions are extant, is another important Pentateuchal Midrash. Additional Midrashic compilations include those to the books of Samuel, Psalms, and Proverbs. Mention should also be made of *Pesiqta* ("Section" or "Cycles") *deRab Kahana* (after a Babylonian *amora*) and *Pesiqta rabbati* ("The Great Cycle"), consisting of homilies on the Torah (Pentateuch) readings that occur on festivals and special Sabbaths.

Haggadic compilations independent of biblical text include *Avot deRabbi Natan, Tanna deve Eliyyahu, Pirqe* ("Chapters") *deRabbi Eliezer,* and tractates *Derekh eretz* ("Correct Conduct"). These primarily deal with ethics, moral teachings, and biblical narrative.

Among the medieval anthologies are the *Yalquṭ* ("Compilation") *Shimoni* (13th century), *Yalquṭ ha-makhiri* (14th century), and *ʿEn Yaʿaqov* ("Eye of Jacob," 16th century). The two most important modern Haggadic anthologies are those of Wilhelm Bacher and Louis Ginzberg.

Codes. The Talmud's dialectic style and organization are not those of a code of laws. Accordingly, codification efforts began shortly after the Talmud's completion. The first known attempt was *Halakhot pesuqot* ("Decided Laws"), ascribed to Yehudai Gaon (8th century). *Halakhot gedolot* ("Great Laws"), by Simeon Kiyyara, followed 100 years later. Both summarize Talmudic Halakhic material, omitting dialectics but preserving Talmudic order and language. The later *geonim* concentrated on particular subjects, such as divorce or vows, introducing the monographic style of codification.

Codification literature gained impetus by the beginning of the 11th century. During the next centuries many compilations appeared in Europe and North Africa. The most notable, following Talmudic order, were the *Hilkhot Harif,* by Isaac Alfasi (11th century), and *Hilkhot Harosh,* by Asher ben Jehiel (13th–14th centuries). Though modelled after *Halakhot gedolot,* the *Hilkhot Harif* encompasses only laws applicable after the destruction of the Temple but includes more particulars. The *Hilkhot Harosh* closely follows Alfasi's code but often also includes the reasoning underlying decisions.

The *Mishne Torah, Sefer ha-ṭurim,* and *Shulḥan ʿarukh*

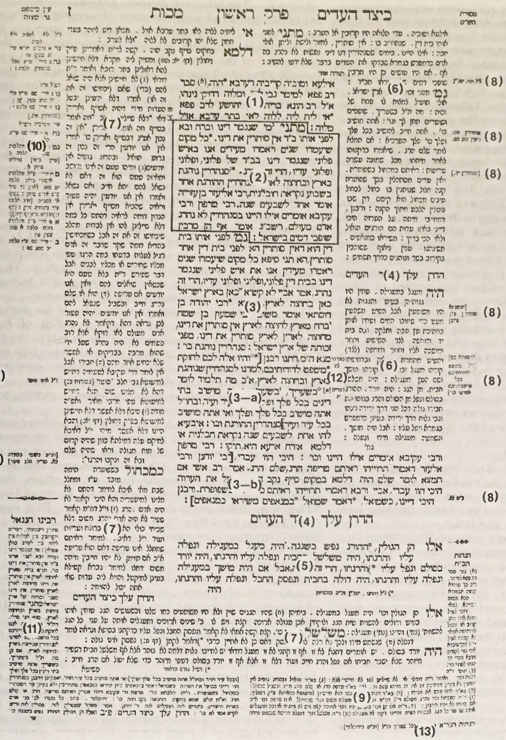
The most important of the topically arranged codifications were: the *Mishne Torah, Sefer ha-ṭurim,* and *Shulḥan ʿarukh.* (1) The *Mishne Torah* ("The Torah Reviewed") by Maimonides (12th century), is a monumental work, original in plan, language, and order; it encompasses all religious subject matter under 14 headings and includes theosophy, theology, and religion. (2) The *Sefer ha-ṭurim* ("Book of Rows," or " Parts"), by Jacob ben Asher (14th century), the son of Asher ben Jehiel, introduced new groupings, dividing subject matter into four major categories (*ṭurim*) reminiscent of the Mishnaic orders; it includes only laws applicable after the destruction of the Temple. (3) The *Shulḥan ʿarukh* ("The Prepared Table") by Joseph Karo (16th century), the last of the great codifiers, is structured after the *Sefer ha-ṭurim,* but presents the Sefardic (Middle Eastern and North African) rather than the Ashkenazic (Franco-German and eastern European) tradition, with decisions largely following those of Alfasi, Maimonides, and Rabbi Asher. When the 16th-century Ashkenazic codifier Moses Isserles added his notes, this became the standard Halakhic code for all Jewry.

Commentaries. The interpretive literature on the Talmud began with the rise of academies in Europe and North Africa. The earliest known European commentary, though ascribed to Gershom ben Judah (10th–11th centuries), is actually an eclectic compilation of notes recorded by students of the Mayence (Mainz) Academy. Compilations of this kind, known as *quntresim* ("notebooks"), also developed in other academies. Their content was masterfully reshaped and reformulated in the renowned 11th-century commentary of Rashi (acronym of *R*abbi *Sh*lomo *Yi*tzḥaqi), in which difficulties likely to be encountered by students are anticipated and detail after detail is clarified until a synthesized, comprehensible whole emerges.

The commentaries of Ḥananel ben Ḥushiel and Nissim ben Jacob ben Nissim, the first to appear in North Africa (11th century), are introductory in nature. They summarize the content of Talmudic discussions, assuming that details will be understood once the general idea becomes comprehensible. This style was later followed by the Spanish school, including Joseph ibn Migash and Maimonides. However, as Rashi's work became known, it displaced all other commentaries. (Note its predominant role in the sample page of Talmud.)

Tosafot, novellae, and *responsa*

A new phase in Talmudic literature was initiated by Rashi's grandchildren, Rabbis Isaac, Samuel, and Jacob, the sons of Meir, who established the school of *tosafot.*


ביצד העדים פרק ראשון מכות ז

הדרן עלך כיצד העדים

Sample page (7a) of the tractate *Makkot* (of the fourth order, *Neziqin*) of the Vilna edition of the Babylonian Talmud, first printed in 1880–86. It discusses the fate of a man who was convicted and escaped and how he is to be judged. Code numbers, a box surrounding the Mishna, and brackets indicating the extent of comments (3-a) and (3-b) have been superimposed onto the original page. (1) End of the Gemara to the previous Mishna. (2) Mishna. (3) Gemara. (3-a) Halakhic Midrash supporting the Mishna. (3-b) Three short comments from Palestine, Sura, and Pumbedita. (4) Mark indicating the end of the chapter. (5) Mishna of the next chapter. (6) Commentary of Rashi (1040–1105). (7) *Tosafot*, discussing special points in the Gemara and Rashi. (8) Cross-reference notes to other Talmudic and Rabbinic sources and textual variants. (9) Notes by Joel Sirkes (1561–1640). (10) References to the codes of Maimonides, Moses of Concy, the *Tur*, and the Shulḥan ʿarukh. (11) Commentary of Hananel of North Africa, early 11th century. (12) References to scripture. (13) Notes by Elijah Gaon of Vilna (1720–97).

By courtesy of the Library of the Jewish Theological Seminary of America, New York; Frank J. Darmstaedter

(These medieval "additions" are not to be confused with the tannaitic Tosefta discussed above.) Reviving Talmudic dialectic, they treated the Talmud in the same way that it had treated the Mishna. They linked apparently unrelated statements from different Talmudic discourses and pointed out the fine distinctions between seemingly interdependent statements. This dialectic style was soon adopted in all European academies. Even the writings of Ravad (Abraham ben David), Zerahiah ha-Levi, and Yeshaya deTrani, three of the most original Talmudists (12th century), reflect the impact of Tosafist dialectic.

The works of Meir Abulafia and Menaḥem Meiri, although of the North African genre, include a strong dialectic element. In Spain such dialectic works were known as *ḥiddushim* or *novellae* (since they sought "new insights"), the most famous being those written by four generations (13th–14th centuries) of teacher and pupil: Ramban (Naḥmanides, or Moses ben Naḥman), Rashba (Solomon ben Adret), Ritba (Yomtov ben Abraham), and Ran (Nissim ben Reuben Gerondi).

A major role in establishing Talmudic authority was also played by the *responsa* literature, replies (*responsa*) to legal and religious questions. Beginning in the 7th century, when the Babylonian *geonim* responded in writing to questions concerning the Talmud, it developed into a branch of Talmudic literature that continued to the present. Then, as now, Talmudic authorities were approached for explanations and decisions. Among the *geonim* the best known were Sherira (10th century) and his son Hai. In the Middle Ages the most important were Alfasi, Ibn Migash (Joseph ibn Migash), Maimonides, Ravad (Abraham ben David of Posquières), Ramban, Rashba, Rosh (Asher ben Jehiel), Ran, and Ribash (Isaac ben Sheshet Perfet).

Writing and printing of the Talmuds. Study in the academies was always oral; hence the question of when the Mishna and Talmud were first committed to writing has been the subject of much discussion. According to some scholars, the process of writing began with Judah ha-Nasi. Others attribute it to the *savoraim.*

The Palestinian Talmud was first printed in Venice (1523–24). All later editions followed this one. Printing of the Babylonian Talmud was begun in Spain about 1482, and there have been more than 100 different editions since. The oldest extant full edition appeared in Venice (1520–23). This became the prototype for later printings, setting the type of page and pagination (a total of close to 5,500 folios). The standard edition was printed in Vilna beginning in 1886. It carries many commentaries and commentaries upon commentaries. In the sample page reproduced here, the Mishna and the Gemara are placed in the centre column of the page and are printed in the heavy type. The commentary of Rashi is always located in the inner column of the page and the *tosafot* in the outer column. Other commentaries and references to legal codes and to scriptural verses surround the major commentaries, in smaller type. Talmudic citations are made by tractate name, folio number, and side of the folio (a or b).

NONLEGAL SUBJECT MATTER

God, Torah, and Israel

Main religious doctrines. While the Talmudic rabbis never formally systematized their beliefs, their underlying religious concepts are clearly reflected in their decisions, ideas, and attitudes. Preeminent in rabbinic thinking were the concepts of God, Torah, and Israel.

God. The rabbinic God was primarily the biblical God who acted in history, the creator and source of life who was experienced through the senses rather than intellect. In reaction to sectarian teachings (*i.e.*, Gnosticism and early Christianity), however, the rabbis stressed God's universality, absolute unity, and direct involvement with the world. His immanence and transcendence (being present in and beyond the universe) were emphasized, and biblical anthropomorphisms (ascribing human attributes to God) were explained metaphorically. The rabbis also stressed an intimacy into the relationship between God and man. God became the father to whom each individual could turn in direct prayer for his needs. To the names YHWH and Elohim, which traditionally were identified with God's mercy and judgment, respectively, the rabbis added new terms reflecting his other attributes—*e.g.*, Shekhina ("Presence"), representing his omnipresence, or immanence; and Maqom ("Place"), his transcendence.

Torah. Torah, in the Talmudic sense, refers to all religious and ethical teachings handed down by tradition. According to the rabbis, God created the Torah long before the world. It contained the eternal divine formula for the world's future workings and thus the answers to all problems for all times and all people. God himself is depicted as studying the Torah, for even he cannot make decisions concerning the world that contradict it.

Israel. The people Israel, according to the rabbis, were chosen by God to be the guardian of his Torah, and, just as God chose Israel, Israel chose God. Thus, the concept of Israel as a nation bound together by an irrevocable commitment to bring the Torah to the world, and bearing corporate responsibility for this mission, was formed. No Jew can free himself from this commitment, but anyone accepting it, regardless of race, becomes a full-fledged Jew with obligations binding him and his descendants.

With this in mind, the rabbis repeatedly emphasized the importance of studying Torah. They pointed out that the Torah is not a declaration of religious beliefs. Rather it is a statement of a discipline regulating each detail of life. Any transgression of this discipline hampers the divine plan of establishing God's way of life in this world.

Sanctification of everyday life

Worship. The intensive rabbinic religious involvement led to the growth of a new concept of worship. While in the Bible worship was usually centred in the sanctuary of the Temple in Jerusalem, the rabbis, particularly after the destruction of the Second Temple (70 CE), attempted to sanctify all of life. Thus, they said that one must bless God upon arising in the morning, before dressing, before and after meals, and in all ordinary daily actions or routines. Each move in life should be an act of worship glorifying God's name.

Messianic kingdom. In rabbinic thinking the establishment of God's kingdom was tied to the Messiah, who was to be a descendant of King David, wise, just, a great scholar, a moral leader, and courageous king. He would redeem the Jews from exile and reestablish their independence in the land of Israel. With this the world would be ushered into a new era of righteousness and universal peace. The rabbis referred to this era as "the world to come," portraying it as an immense academy in which the righteous would study Torah without interruption. They refrained from describing it further, saying that human language and fantasy are inadequate to its wonders.

The nature of the Messiah and the time of his arrival raised much speculation. Following the defeat of Bar Kokhba, leader of the revolt against Roman rule (135 CE), the Messiah's coming, in rabbinic thought, faded into the mysterious and distant future, and descriptions concerning his personality assumed supernatural overtones.

For a fuller discussion of major religious doctrines, see below, *Basic beliefs and doctrines.*

Doctrine of man. The fate of man, his achievements and failures, his being and nothingness, occupy an important place in Talmudic literature. The rabbis' concept of man was a universal one. While they assumed that Jews are bound by greater religious duties than others, they considered all men equal, all created in the image of God. "Therefore, but a single man was created . . . That none should say to his fellow, 'My father was greater than thy father'" (tractate *Sanhedrin*).

The world, according to the Talmud, was created for the sake of man, and it is incumbent upon him to keep it in order. His responsibility begins at home. Man must care for his health, marry, build a family, provide for and educate children, honour parents, friends, and elders. He also carries social responsibilities and has to be part of the community. He must learn a trade and work so that he does not become a burden to the community.

The uniqueness of man in this world, likened by the Talmud to the uniqueness of God in the universe, lies in his freedom of choice. Nature follows its laws and angels their missions, but man is his own master. In contrast to St. Paul's doctrine that the original sin of Adam made sin an integral part of human nature, the rabbis considered man a wondrous and harmonious being. The duality of his nature was explained by the existence of a good and bad impulse, personified by two angels, *yetzer ha-ṭov* (the good inclination) and *yetzer ha-raʿ* (the evil inclination), which enter each man after birth. It is the duty of man to overcome his evil inclination, and it is for this that he is rewarded. Moreover, since there is corporate responsibility, not only is the sinner punished but the community at large also suffers. Here again, however, man is his own master. He can reverse the course of sin and punishment by repentance. Although repentance may be accompanied by formal and ceremonial acts, such as fasting, its basic principle is the renunciation of the sin and the wholehearted decision not to repeat it. When a man transgresses against God, his sin is forgiven by repentance alone, but, when he transgresses against his fellow man, he must make good his wrongdoing as well as repent.

The *yetzer ha-ṭov* and *yetzer ha-raʿ*

Medicine and science. The Talmud devoted considerable attention to the maintenance of good health, regarding it a religious duty. A keen understanding of the importance of hygiene in preventing illness was reflected in an emphasis upon bodily cleanliness. The rabbis also stressed the necessity for moderation in eating and drinking and the importance of a proper diet. The Talmud prescribed remedies for illnesses and mentioned surgical techniques, such as cesarean section.

Religious concerns surrounding the calendar, prohibitions against planting seeds of different kinds together, dietary laws, and Sabbath-walking limits resulted in an intense rabbinical interest in astronomy, zoology, mathematics, and geometry.

Legend and folklore. Side by side with the Midrashic Haggada, which was the outgrowth of Bible exegesis and developed in the academies, the Talmuds and Midrashic collections contain a large quantity of Haggadic material with mythological rudiments, allusions to pagan beliefs and customs, and folkloristic elements of a world strange to the rabbis. Folktales and legends, animal lore, and adventure narratives, containing pagan ideas and beliefs, that were told by their Gentile neighbours were no doubt a major attraction to the common Jews, especially those in the countryside (the *'am ha-aretz*, or "people of the land"). The rabbis realized the great danger involved in this situation and developed their own folk material. They adopted the dramatic and artistic parts of these stories but rejected the unwanted elements, replacing them with their own ideas. Thus the animals and birds in fables quote the Bible and discuss it in the same manner that the rabbis do.

The challenge of pagan myths, legend, and folklore

Ancient mythology seems to have been well known and liked by the Jewish masses. Again, in order to fight its influence, the rabbis reworked its content in their own spirit. They retained the mythological suspense—the sea tries to drown the earth—but there is no mythological struggle between equal powers; angels try to prevent the creation of man, but they do not possess titanic power. All are subdued by the command of God. Thus, the rabbis transformed the ancient myths into dramatic evidence against polytheism. (See also below, *Jewish myth and legend.*)

Astrology, magic, and divination. Astrology was a recognized science in the ancient world. The rabbis could not reject it entirely, and some concluded that the power of the stars is confined to Gentiles. Others made it part of God's order, saying that stars influence this world in the same way that climate influences plants. The rabbis strenuously objected to omens and other forms of divination because they considered them magic. Dreams were considered by some rabbis as meaningless, while others saw in them an element of prophecy.

The rabbis believed in the efficacy of magic but strenuously objected to its practice. They permitted only magic that had been proved effective in healing. They also permitted the use of incantations for the purpose of counteracting the hold of magic. Because of their supposedly protective nature, the use of amulets was also countenanced.

Demons (evil spirits) and the evil eye

The existence of a demonic kingdom was accepted by the rabbis without question. Evil spirits are invisible and fill the nether world. They avoid sunlight and concentrate in waters and deserted places. They also mingle with people, trouble them, and help them. They have passions and are born and die like people. However, they also have some of the traits and powers of angels. The evil eye was considered as dangerous as evil spirits. It was thought that for mysterious reasons some people have the power to injure others by looking at them and that it is generally jealousy that triggers this effect. The rabbis, however, repeatedly emphasized that all of these strange powers are under the divine government and, moreover, that they cannot hurt the pious.

TALMUDIC LAW AND JURISPRUDENCE

Sacral basis of Talmudic law

Unlike the Romans, who considered ritual law (*fas*) God-given and social law (*lex*) man-made, the rabbis believed all Jewish law to be of divine origin. Thus, for example, unfairness in labour relations was considered a religious sin and caring for the sick a religious obligation. Though familiar with the concept of natural law (ethical principles inherent in the nature of things and apprehensible through human reason), the rabbis objected to making nature the basis of law. Even rabbinic ordinances were regarded as having validity only because the authority of the rabbis is sanctioned by the Torah.

Methods of arriving at legal principle and decisions. Ancient Halakha knew no controversy. The earliest controversy dates to the pre-tannaitic *zugot.* Hillel and Shammai differed on significant issues, and, with the rise of their schools, Halakhic uniformity began to crumble. Halakha became a scholastic discipline that developed in academic rather than judicial settings, more and more issues remaining unresolved. Over 300 controversies between the schools of Hillel and Shammai (called the House of Hillel and the House of Shammai, respectively) are reported in Talmudic sources. As time passed, disputes proliferated even more and were considered legitimate provided they conformed to the rule of Halakhic discipline.

No attempt was made to restore Halakhic uniformity until the beginning of the 2nd century CE. Controversies were sometimes resolved by citing old traditions, by establishing precedents, or, when the sages could convene, by vote taking.

At Yavne, Gamaliel II, the president of the revived Sanhedrin (*c.* 80–*c.* 115 CE), attempted to suppress diversity of opinion, but failed. The right to differ was already established. Moreover, in the Halakhic collection compiled at Yavne (tractate *'Eduyyot*), the views of individual scholars were preserved. The sages at Yavne, however, did take a major step toward restoring Halakhic consistency by upholding the generally more lenient views of the House of Hillel over those of the House of Shammai, thus establishing the Hillelite tradition as the main trend of rabbinic Judaism.

The principle that differing opinions should be recorded was followed by Judah ha-Nasi in his Mishna. Modern scholars differ as to whether he meant to compile a code of law or merely a Halakhic collection. The *amoraim,* however, accepted his Mishna as the definitive code and introduced a set of guidelines according to which disputes were decided. Thus, for example, collective (". . . the sages said") and individual opinions stated anonymously were taken as law; Akiba's decisions were upheld over those of his colleagues. Similar guidelines developed also with regard to amoraic controversies.

Guidelines for settling Halakhic controversies

With the completion of the Talmud, a new phase in Halakhic development began. Not only were there two different Talmuds and a large Haggadic literature but even within each of the Talmuds diversified opinions were reported. The *geonim* laid down rules governing the use of this enormous literature for lawmaking. They designated the Babylonian Talmud the highest authority, taking the Palestinian Talmud into consideration only when it did not disagree with the Babylonian or when the latter expressed no opinion on a subject. They also deprived the Haggadic literature of Halakhic authority and set guidelines for the precedence of opinion among *amoraim.* These geonic rules served as the basis of all future codifications.

After the geonic period two methods of decision making were applied. The first of these relied primarily upon the authoritative codes. The Mediterranean rabbis, for example, made the code of Maimonides the source of all of their lawmaking. The second method relied on the original Talmudic sources for decision making. This method was applied by the Tosafists and their followers, who, though they consulted the older codes, did not accept them as the final authorities. The *responsa* literature represents a synthesis of these two methods. Although it makes use of codes as the main source of law, its decisions are always accompanied by a discussion and analysis of earlier relevant literature. This approach has been used by rabbis to the present day.

In addition to the above, in particular instances throughout the ages rabbinic authorities promulgated ordinances (*taqqanot*) and edicts (*gezerot*). These were made in response to pressing needs of time and circumstance, and this form of lawmaking was most frequently used by rabbinic synods in the Middle Ages.

The Great and Lesser Sanhedrins

Administration of justice. *Courts.* A comprehensive judicial system is described in Talmudic sources. The highest court was the Great Sanhedrin. It consisted of 71 members and convened daily in one of the Temple halls. It was the highest legal and religious authority in the country and had exclusive jurisdiction over matters of a national and public nature. It also functioned as the court of appeals, dealing with cases that were not resolved by the lower courts.

Next in line of judicial authority was the Lesser Sanhedrin. Each town with a population of 120 or more had a court of this kind. These courts each consisted of 23 members and dealt with cases involving capital punishment.

The members of the Sanhedrins had to be ordained, pious, mature in age, sound in mind and body, of wide knowledge, and of pure Jewish descent. Persons who were too old or who had never had children were ineligible, for it was thought that they might not be merciful.

The lower courts dealt with all remaining cases. Each consisted of three members and convened on Mondays and Thursdays. In cases involving a penalty the three judges had to be ordained, but in those involving ordinary monetary litigation ordination was not required. In the latter type of case, concerned parties were allowed the alternative of setting up ad hoc arbitration bodies.

Rules of evidence. Jewish law was extremely strict regarding evidence acceptable in court. In cases entailing physical punishment, no circumstantial evidence, confession, or self-incrimination was recognized. The testimony of two eyewitnesses who confronted the defendant was required. In monetary cases documentary evidence and, at times, oaths were acceptable. Any mental or moral defects or self-interest in the case disqualified witnesses. Relatives could not serve as judges or witnesses.

Trial procedure. Jewish law knows of no lawyers. After the facts were presented, the court investigated, deliberated, and made its decision by voting. Both sides had to be treated equally, even to the point of seeing to it that neither should be dressed more richly than the other. Each side could be heard only in the presence of the other.

In the trial procedure of capital cases, there was a clear tendency toward bias in favour of the defendant. Thus, only the judges could argue for conviction, but all present could argue for acquittal. The most junior judges voted first so that they would not be unduly influenced by their seniors. A majority of one was sufficient for acquittal, but a majority of two was necessary for conviction. A verdict of acquittal could be reached on the same day but one of conviction only on the following day. When the court erred, only its convictions, and not its acquittals, were reversed.

Death penalties

Criminal law. In Jewish law, ritual and nonritual transgressions were crimes punishable by court. Each of the 36 most severe transgressions (*e.g.*, adultery, sodomy, idolatry, sorcery, or murder) carried one of four types of death penalty (stoning, burning, beheading, and strangling). Rabbinic law, however, tended to minimize the practice of capital punishment. The rigorous cross-examination of witnesses and the warning of impending punishment that the transgressor had to receive immediately before committing his crime made it almost impossible to reach a death verdict.

If despite all of this a death verdict was reached, every legal effort was made to allow for a last-minute reversal. Execution was expedited and carried out in the most humane manner possible, the accused being given an opiate before dying. To show their compassion the judges fasted on the day of execution. According to tradition the death penalty was abolished 40 years before the destruction of the Temple, when the Great Sanhedrin was exiled from the Temple complex.

The punishment for 207 other transgressions (*e.g.*, perjury, some forms of incest, the eating of forbidden foods) was flagellation. Here, too, the rabbis tended to be lenient. As in capital cases, a rigorous cross-examination and a warning were required. The maximum number of stripes administered was 39. Prior to flagellation the transgressor was examined medically to determine the number of stripes he could withstand.

Side by side with the above penalties, the courts also inflicted *makkat mardut* (disciplinary stripes) and excommunication in cases where regular flagellation could not legally be applied. These two punishments were generally used in Babylonia, where ordained courts did not exist. It should be mentioned also that the Mishna includes a few obscure references to a form of imprisonment used instead of capital punishment.

Civil and social law. Although the rabbis considered both ritual and nonritual law sacred, they demonstrated great independence in supplementing the relatively brief relevant scriptural comments and regulations with a comprehensive system of civil and social law. In response to variations in social and economic circumstances, certain differences in Palestinian and Babylonian Talmudic law emerged. The Babylonian rabbis, for example, recognized the law of the state as binding in monetary matters, while the Palestinian rabbis did not. In general, however, Jewish civil law developed relatively autonomously. In instances where the rabbis did adopt alien legal concepts, they elaborated upon them until they could be fully integrated into the spirit and structure of Jewish law.

Social welfare, torts, family law, financial law

The following are some of the areas covered: (1) Social welfare: a comprehensive social welfare system was worked out, including obligations to provide for children, educate them, and train them for a profession. Regulations of charity, medical assistance, and burial of the dead were established. (2) Torts: included were all damages caused by a person directly or indirectly via his property. The main aim was to compensate for damages. Consequently, no torts were classified as criminal. Even "an eye for an eye" was interpreted to mean financial compensation. (3) Family law: included were regulations concerning marriage and divorce procedures and the innovation of the *ketubba* (marriage contract), which spells out the mutual obligations of husband and wife in the areas of finance, medical care, clothing, housework, sexual relations, and child care. According to biblical law, the right to inherit belongs to sons first. To protect the rights of wives and daughters, rabbinic law obligated the sons to maintain the widows and unmarried daughters. (4) Financial law: except for Gen. 23:9 ff., Jer. 32:10, and Ruth 4:8, Scripture makes no reference to transaction procedures. The growth of finances, industry, and land estates led the rabbis to develop laws concerning contracts, partnerships, and legal arrangements to circumvent the biblical prohibition against usury. A series of modes of transaction effecting the transfer and acquisition of property evolved. Labour relations, rents, and leases were also carefully regulated.

THE TALMUD TODAY

Role in the Jewish community. With the rebirth of a Jewish national state (since 1948) and the concomitant revival of Jewish culture, the Talmud has achieved renewed importance. Orthodox Jewry has always focussed upon its study and has believed it to be the absolute Halakhic authority. This belief has now become even further intensified. While rabbinic courts in Israel have jurisdiction only in the area of family life, it has become one of the aims of religious (Orthodox) Jewry there to establish Talmudic law as the general law of the state.

It should also be noted that, aside from the special case of Israel, the legal system described above has continued to function down to the present day in Jewish communities all over the world. The jurisdiction of rabbinic courts is voluntarily accepted by Orthodox Jews. These courts continue to exert authority, especially in the areas of family and dietary law, the synagogue, and the organization of charity and social activity.

Conservative Jewry, too, has always been committed to rabbinic tradition. It has, however, conceptualized this tradition as an evolutionary process in which Halakha changes to meet the challenge of new conditions. Professional scholarship was considered crucial for understanding the furthering of this process. More recently, however, as a result of revived nationalism, new emphasis has been put upon lay education. Thus, a network of day schools and higher institutions of learning in which rabbinic tradition occupies a major role in the curriculum has been established. Scores of young Conservative Jews now search in the Talmud for answers to crucial problems, such as abortion and civil violence.

Classical (19th-century) Reform Judaism not only dissociated itself from the Talmud but negated it. More recently, however, Reform leaders have been inclined to reestablish some measure of ritual practice and rabbinic climate. Thus, it is now not unusual to find them stating their decisions in the form of *responsa* and using the rabbinic style of argument and even the casuistic type of Talmudic dialectic (*pilpul*) to justify their religious practices.

Talmudic scholarship. Although Talmudic scholarship continues to be advanced by individuals in a number of countries, its two main centres are in Israel and the United States. The Israeli centre has tended to focus upon research of a critical nature. Like Bible criticism, this work is divided between source criticism (*i.e.*, discovering the different sources, their dates, and the methods by which Talmudic literature was formed) and textual criticism (*i.e.*, establishing the correct text and reading). Research is also being done on Haggadic concepts and thinking, Talmudic law, and Halakhic development.

Talmudic scholarship in the United States has tended to be more philosophically and historically oriented. There has been great interest in the development of Halakha and in folklore and custom. Essential work has been done and continues to be done in the areas of source criticism. A work unique in scope and method is S. Lieberman's commentary on the Tosefta. (H.Z.D./L.H.S./Ed.)

Basic beliefs and doctrines

Judaism is not and cannot be viewed as an abstract intellectual system, although some of its affirmations may be couched in such terms. It affirms divine sovereignty disclosed in creation (nature) and in history, without necessarily insisting upon—but at the same time not rejecting—metaphysical speculation about the divine (see below *Jewish philosophy*). It insists that the community has been confronted by the divine not as abstraction but as person, with whom the community and its members enter into relationship. It is—as the concept Torah indicates—a program of human action, rooted in this personal confrontation. Further, the response of this particular people to its encounter with God is viewed as significant for all mankind. The community is called upon to express its loyalty to God and the Covenant by exhibiting solidarity within its corporate life on every level—including every aspect of human behaviour, from the most public to the most private. Thus, even Jewish worship is communal celebration of the meetings with God in history and in nature. Yet the particular existence of the Covenant people is not thought of as contradicting but rather as enhancing human solidarity. This people, together with all men, is called upon to create political, economic, and social forms that will affirm divine sovereignty—embody it in communal existence. This task is carried out in the belief not that man will succeed solely by his own efforts in these endeavours but that these sought-after human relationships have both their source and their goal in God—who assures their actualization. Within the sphere of his existence in the community, each Jew is called upon to realize the Covenant in his personal intention and behaviour.

The Judaic and the human

In considering the basic affirmations of Judaism from this point of view it is best to allow indigenous formulations rather than systematic statements borrowed from other traditions to govern the presentation.

GOD

An early statement of basic beliefs and doctrines emerged in the liturgy of the synagogue some time during the last pre-Christian and first Christian centuries, although there is evidence that such formulations were not absent from the Temple cult that came to an end in the year 70 CE. A section of the *Siddur* (order of worship, or prayerbook) that has as its focus the recitation of a series of biblical passages (Deut. 6:4–9; Deut. 11:13–21; Num. 15:37–41) takes its name from the first of these, Shema ("Hear"): "Hear, O Israel! the Lord is our God, the Lord alone" (or ". . . the Lord our God, the Lord is one"). In the Shema—often regarded as the Jewish confession of faith, or creed—the biblical material and accompanying benedictions are arranged to provide a unified statement about God and his relationship to the world and Israel, as well as Israel's obligations toward and response to God. In this statement, God, who is the Creator of the universe and who has chosen Israel in love ("Blessed art thou, O Lord, who has chosen thy people Israel in love"), expressed by the giving of Torah, is declared to be "one"; his love is to be reciprocated by men who lovingly obey Torah and whose obedience is rewarded and rebellion punished. The goal of this obedience is God's "redemption" of Israel, a role foreshadowed by his action in bringing Israel out of Egypt.

Unity and uniqueness. At the centre of this liturgical formulation of belief is the concept of the divine unity. In its original setting, it may have served as the theological statement of the reform under Josiah, king of Judah, in the 7th century BCE when worship was centred exclusively in Jerusalem, and all other cultic centres were rejected, so that the existence of one shrine only was understood as affirming one deity. The idea, however, acquired further meaning. It was understood toward the end of the pre-Christian era to proclaim—over against dualistic religious formulations in the Greco-Roman world—the unity of divine love and divine justice, as expressed in the divine names YHWH and Elohim, respectively. A further expansion of this affirmation is found in the first two benedictions of this liturgical section, which together proclaim that the God who is the Creator of the universe and the God who is Israel's ruler and lawgiver are one and the same—as over against religious positions that insisted that the Creator God and the lawgiver God were separate and even inimical. Subsequently, this affirmation was developed in philosophical and mystical terms by both medieval and modern thinkers.

Divine love and justice

Creativity. As has been noted, this "creed," or "confession of faith," underscores in the first benediction the relation of God to the world as that of Creator to creation. "Blessed art thou, O Lord our God, King of the Universe, who forms light and creates darkness, who makes peace and creates all things." It adds the assertion that his activity is not in the past but is ongoing and continuous, for "he makes new continually, each day, the work of creation"; thus, unlike the deity of the Stoic world view, he remains actively present in nature. This "creed" is concerned as well to come to terms with the ever-present problem of evil. Paraphrasing Isa. 45:7, "I form the light and create darkness; I make peace, and create evil," it changes the last word to "all" (or "all things") rather than "evil." The change was clearly made to avoid the implication that God is the source of moral evil. Judaism, however, did not flinch from confronting the problem of pain and suffering in the world and affirming the paradox of suffering and divine sovereignty, of pain and divine providence, refusing to accept the concept of a partial God—a God that is Lord over only the harmonious and pleasant aspects of reality.

Activity in the world. The second and the third benedictions deal with divine activity within the realm of history and human life. God is teacher of men through the giving of instruction (Torah; see above); he acts in the life of mankind in historical events; he has chosen a particular people—Israel—in love to witness to his presence and his desire for a perfected society; he will, as redeemer, enable man to experience that perfection. These activities, together with creation itself, are understood to express divine compassion and kindness as well as justice (judgment), recognizing the sometimes paradoxical relation between them. Taken together, they disclose Divine Providence—God's continual activity in the world. The constant renewal of creation (nature) is itself an act of compassion overriding strict justice and affording mankind further opportunity to fulfill the divinely appointed obligation. Yet the basically moral nature of God is asserted in the second of the biblical passages that form the core of this liturgical statement (Deut. 11:13–21). Here, in the language of its agricultural setting, the community is promised reward for obedience and punishment for disobedience. The intention of the passage is clear: obedience is rewarded by the preservation of order, so that the community and its members find wholeness in life; while disobedience—rebellion against divine sovereignty—shatters order, so that the community is overwhelmed by adversity. The passage of time has made the original language unsatisfactory (promising rain, crops, and fat cattle), but the basic principle remains, affirming that, however difficult it is to recognize the fact, there is a divine law and judge. Support for this affirmation is drawn from the third biblical passage (Num. 15:37–41), which explains that the fringes the

Divine Providence

Israelites are commanded to wear on the corners of their garments are reminders to observe the commandments of God, who brought forth Israel from Egyptian bondage. The theme of divine redemption is elaborated in the concluding benediction to point toward a future in which the as yet fragmentary rule of God will be brought to completion: "Blessed is his name whose glorious kingdom is for ever and ever."

Otherness and nearness. Within this complex of ideas, other themes are interwoven. In the concept of the divine Creator there is a somewhat impersonal or remote quality—of a power above and apart from the world—which is underscored by such expressions as the trifold declaration of God's holiness, or divine otherness, in Isaiah 6:3: "Holy, holy, holy is the Lord of hosts . . ." The development of surrogate divine names for biblical usage, as well as the substitution of Adonai ("my Lord") for the tetragrammaton (YHWH) in the reading of the Bible itself, suggests an acute awareness of the otherness of God. Yet this has as its countertheme the affirmation of divine nearness. In the biblical narrative it is God himself who is the directly active participant in events, an idea that is emphasized in the liturgical narrative (Haggada) recited during the Passover meal (seder): "and the Lord brought us forth out of Egypt—not by an angel, and not by a seraph, and not by a messenger . . . ." The surrogate divine name *Shekhina,* the Present One, is derived from a Hebrew root meaning "to dwell," again calling attention to divine nearness ("presence"). The relationship between these two affirmations, otherness and nearness, is expressed in a Midrashic statement, "in every place that divine awesome majesty is mentioned in Scripture, divine abasement is spoken of, too."

The divine "thou"

Closely connected with these ideas is that of divine personhood, most particularly disclosed in the use of the pronoun "thou" in direct address to God. The community and the individual, confronted by the Creator, teacher, redeemer, addresses the divine as living person, not as theological abstraction. The basic liturgical form, the *berakha* ("blessing"), is usually couched in the second person singular: "Blessed art thou . . . ." This relationship, through which remoteness is overcome and presentness is established, illuminates creation, Torah, and redemption, for it reveals the meaning of love. From it flow the various possibilities of expressing the divine–human relationship in personal, intimate language. Sometimes, especially in mystical thought, such language becomes extravagant, foreshadowed by such vivid biblical metaphors as the husband–wife relation in Hosea; the "adoption" motif in Ezek. 16; and the firstborn-son relation (Ex. 4:22). Nonetheless, although terms of personal intimacy are used widely to express Israel's and man's relationship with God, such usage is restrained by the accompanying sense of divine otherness. This is to be seen in the liturgical "blessings," where, following the direct address to God, in which the second person singular pronoun is used, the verbs, with great regularity, are in the third person singular, thus providing the requisite tension between nearness and otherness, between the impersonal and the personal.

Modern views of God. The Judaic affirmations about God have not always been given the same emphasis nor have they been understood in the same way. This was true in the Middle Ages, among both philosophers and mystics, as well as in modern times. In the 19th century, western European Jewish thinkers attempted to express and transform these affirmations in terms of German Idealist philosophy: more recently, philosophical Naturalism was offered as the suitable content of Judaism, while still retaining the traditional God language. The meaningfulness of the whole body of such affirmations, moreover, has been called into question by the philosophical schools of Logical Positivism and Linguistic Analysis. Most recently, the destruction of 6,000,000 Jews during the Nazi period has raised the issue of the validity of such concepts as God's presence in history, divine redemption, the covenant, and the chosen people. In every case, however, it is with the structure of ideas here noted that these challenges must deal.

ISRAEL (THE JEWISH PEOPLE)

Choice and covenant. The concluding phrase of the second benediction of the liturgical section referred to above reads: "who has chosen thy people Israel in love." Here the basis of the relationship between God and Israel set forth in the biblical narrative is clearly and succinctly stated: the choice of this people was determined by no other factor than divine love. The patriarchal narratives, beginning with the 12th chapter of Genesis, presuppose the choice, which is set forth explicitly in Deut. 7:6–8 in the New Jewish Version:

> For you are a people consecrated to the Lord your God: of all the peoples on earth the Lord your God chose you to be His treasured people. It is not because you are the most numerous of peoples that the Lord set His heart on you and chose you—indeed you are the smallest of peoples; but it was because the Lord loved you and kept the oath He made with your fathers that the Lord freed you with a mighty hand and rescued you from the house of bondage, from the power of Pharaoh king of Egypt.

Later rabbinic traditions on occasion sought to base the choice upon some special merit of Israel, and the medieval poet and theologian Judah ha-Levi suggested that the openness to divine influence originally present in Adam continued only within the people of Israel.

However understood, the background of this choice is the recurring disobedience of mankind narrated in Genesis 2–11. Abraham and his descendants are singled out not merely as the object of the divine blessing but also as its channel to all mankind. The choice, however, demands a reciprocal response from Abraham and his lineage. That response is obedience, as exemplified in the first instance by Abraham's readiness to leave his "native land" and "father's house" (Gen. 12:1). This twofold relationship was formalized in a mutually binding agreement, a covenant between the two parties. The covenant, thought by some modern biblical scholars to reflect the form of ancient suzerainty treaties, indicates (as in the Ten Utterances) the source of Israel's obligation—the acts of God in history—and the specific requirements those acts impose. The formalization of this relationship was accomplished by certain cultic acts that may, according to some contemporary scholars, have been reenacted on a regular basis at various sacred sites in the land, eventually being centralized in Jerusalem. The content of the covenantal obligations thus formalized was Torah. Israel was bound in obedience, and Israel's failure to obey provided the occasions for the prophetic messages. The prophets, as spokesmen for God, called the community to renewed obedience, threatened and promised disaster if such was not forthcoming, and—recalling the source of the choice in divine love—sought to explain its persistence even when, strictly understood, the covenant should have been repudiated by God.

The binding agreement between God and Israel

The choice of Israel has its concrete expression in the requirements of the precepts (*mitzwot,* singular *mitzwa*) that are part of Torah. The blessing recited before the public reading of the pentateuchal portions on Sabbath, festivals, holy days, fasts, and certain weekdays refers to God as "He who chose us from among all the peoples and gave us His Torah," thus emphasizing the intimate relationship between the elective and revelatory aspects of God.

Priest and witness to the nations

Israel's role was not defined solely in terms of its own obedience to the commandments. As noted, Abraham and his descendants were seen as the means by which the estrangement of disobedient mankind from God was to be overcome. Torah was the formative principle underlying the community's fulfillment of this obligation. Israel was to be "a kingdom of priests and a holy nation" (Ex. 19:6) functioning within mankind and for its sake. This task is enunciated with particular earnestness in the writings of the prophets. In Isa. 43–44, Israel is declared to be God's witness and servant who is to bring the knowledge of God to the nations. In chapter 42 of the same book Israel is declared to be a "covenant of the people, a light to the nations, to open the blind eyes, to bring out the prisoners from the prisons, and them that sit in darkness out of the prison house" (42:6–7). This double motif of a chosen people and witness to the nations, joined to that of the righteous king, developed in the biblical and postbiblical periods into messianism in its several varieties (see below *Eschatology*).

The intimate relation between choice, covenant, and Torah determined the modality of Israel's existence. Religious faith, far from being restricted to or encapsulated in the cult, found its expression in the totality of communal and individual life. The obligation of the people was to be the true community, in which the relationship between its members was open, in which social distance was repudiated, and in which response to the divine will expressed in Torah was called for equally from all. One of the important recurring themes of the prophetic movement was the adamant rejection of any tendency to limit divine sovereignty to the partial area of "religion," understood as the realm of the priesthood and cult. Subsequent developments continued this theme, although it appeared in a number of other forms. Pharisaic Judaism and its continuation, rabbinic Judaism, down to modern times has resolutely held to the idea of the all-pervasive functioning of Torah, so that however the various Jewish communities over the centuries may have failed to fulfill the ideal, the self-image of the people was that of "holy community."

Israel and the nations. The double motif of "treasured people" and "witness" was not without its tensions as it functioned in ongoing history. Tensions are especially visible in the period following the return from the Babylonian Exile at the end of the 6th and beginning of the 5th centuries BCE. It is, however, doubtful whether the use of such terms as nationalism, particularism, or exclusivism (as opposed to universalism) are of any great help in understanding the situation. Emphasis has, for example, been laid upon Ezra 9:2 and 10:2, in which the reestablished community is commanded to give up wives taken from "the peoples of the land." This is taken as indication of the narrow, exclusivistic, nationalistic nature of Judaism, without reference to the situation in which a harassed contingent of returned exiles sought to maintain itself in a territory surrounded by politically unfriendly if not hostile neighbours. Nor does this represent racialism, since foreigners were admitted to the Jewish community, and in the following centuries some groups engaged in extensive missionary activities, appealing to the individuals of the nations surrounding them to join themselves to the God of Israel, who was the one true God, the Creator of the heavens and earth.

The welding of exclusivism and universalism

A more balanced view recognizes that within the Jewish community religious universalism was affirmed at the same time and by the same people who understood the nature of Jewish existence in politically particularistic (*i.e.*, nationalistic) terms. To neglect either side is to distort the picture. In no case was the universalism disengaged from the reality of the existing community, even when it was expressed in terms of the ultimate fulfillment of the divine purpose, the restoration of the true covenantal relationship between God and all mankind. Nor was political particularism, even under circumstances of great provocation and resentment, misanthropic. The most satisfactory figure in describing the situation of the restored community, and one that continues to be useful in dealing with later episodes, is that of the human heartbeat, made up of two functions, the systole, or contraction, and the diastole, or expansion. There have been several periods of contraction and of expansion throughout the history of Judaism. The emphasis within the abiding tension has been determined by the historical situation in which the community has found itself. To generalize in one direction or the other is fatal to an understanding of the history and faith of the "holy community."

The people and the land. Closely related to the concept of Israel as the chosen, or Covenant, people is the role of the land of Israel. In the patriarchal stories, settlement in Canaan is an integral part of the fulfillment, from the divine side, of the Covenant. The goal of the Israelites escaping from Egypt is the same land, and entry into it is understood in the same fashion. The return from the Babylonian Exile, too, is seen in the same light. As there was the choice of a people, so was there the choice of a land—and for much the same reason. It was to provide the setting in which the community could come into being as it carried out the divine commandments. This choice of the land contrasts significantly with the predominant ideas of other peoples in the ancient world, in which the deity or divinities were usually bound to a particular parcel of ground outside of which they lost their effectiveness or reality. Though some such concepts may very well have crept into Israelite thought during the period of the kings (from Saul to Jehoiachin), the crisis of the Babylonian Exile was met by a renewal of the affirmation that the God of Israel was, as Lord of all the earth, free from territorial restraint, although He had chosen a particular territory for this chosen people. Here again the twofold nature of Jewish thought becomes apparent, and both sides are to be affirmed or the view is distorted. Following the two revolts against Rome (66–73 CE and 132–135 CE), the Jews of the ever-widening dispersion continued, as they had before these disasters, to cherish the land. Once again it became the symbol of fulfillment, so that return to it was looked upon as an integral part of messianic restoration.

The land as ideal and reality

The liturgical patterns of the community, insofar as they were concerned with natural phenomena (*e.g.*, planting, rainfall, harvest, and the annual cycle) rather than historical events, were based on geography, topography, and agricultural practices of the land, viewed as paradigmatic. Although Jews continued to live in the land, yet for most in the distant dispersion it was idealized and viewed primarily in eschatological terms—at the end of days, in the world to come. The 11th-century poet Judah ha-Levi not only longed for it in verse but also gave it a significant role in his theological interpretation of Judaism and eventually sought to return to it from his native Spain. It was not, however, until the 19th century that the land began to play a role other than the goal of pilgrimage or of occasional settlement by pietists and mystics. At the end of that century the power of the utopian concept was released in eastern Europe in a cultural renaissance that focussed, in part, on a return to the land and, in western and central Europe, in a political movement coloured by nationalist motifs in European thought. The coming together of these two gave rise to Zionism. The political movement reflected a dissatisfaction with the view of the Jews as merely a body or organization of religious believers—like the Christian churches—an interpretation that had become dominant following the political emancipation of the Jews in the period after Napoleon. The political emphasis of Zionism aroused considerable opposition from those Jews who were convinced of the necessity of a churchly definition of Judaism parallel to the Roman Catholic and Protestant communions. While this conflict erupted in bitter debate during the first half of the 20th century, the events of the Nazi period in Europe brought it to a close, except for some sporadic renewals on the part of numerically insignificant groups. For the most part, although there are few satisfactory formulations—theological or secular—there is a working consensus that acknowledges a significant role to the land and recognizes that a churchly definition of the Jewish community, while strategically acceptable in some situations, does not do justice to history and is not theologically sound if it suggests that Judaism merely consists of abstract doctrines and dogmas. Some Jews, however, argue that whatever the past has been, the future of the Jewish community is with those movements in the modern world that cut across or transcend the particularity Zionism represents.

Modern views of the people Israel. The nature of the people Israel and of the land of Israel has been variously interpreted in the history of Jewish thought. In modern times, some interpretations have been deeply influenced by contemporary political and social discussions in the general community. Thus, for example, Zionist theoreticians were influenced by concepts of political nationalism on the one hand and by socialist ideas on the other. Further, the challenge to traditional theological concepts in the 19th century raised issues about the meaning of the choice of Israel, and Jewish thinkers borrowed from romantic nationalism such ideas as the "genius" of the people.

Modern reinterpretations or rejections of chosenness

Most recently, attempts have been made to approach the question sociologically, dismissing the theological mode as unhelpful. The concept of the chosen people is then understood to indicate a specific role deliberately undertaken by the Jewish people and similar to that espoused by other groups (*e.g.*, "Manifest Destiny" by the American people). The establishment of the State of Israel has motivated some thinkers to call for a repudiation of the idea, in keeping with the position that normal existence for the Jews requires the dismissal of such concepts. Although only a small minority of Jewish thinkers espouse this position, the doctrine of the choice is not without its theological difficulties even for those who continue to affirm it.

MAN

The image of God. In Gen. 1:26, 27; 5:1; and 9:6 two terms occur, "image" and "likeness," that seem to indicate clearly the biblical understanding of man's essential nature: he is created in the image and likeness of God. Yet the texts in which they are used are not entirely unambiguous; the idea they point to does not appear elsewhere in Scriptures; and the concept is skirted cautiously in the rabbinic interpretations. What the image and likeness of God or the divine image refer to in the biblical text is not made explicit, and, in the light of the psychosomatic unity of man that dominates the biblical concepts, it is not possible to escape entirely from the implication of "bodily" similarity. What the terms meant in their context at the time and whether they reflect mythological usages taken over from other Middle Eastern thought is a question that is by no means answered. Evidence of the problematic nature of the concept is found in rabbinic Judaism. Akiba (2nd century CE) ignored the usages in Gen. 1 and 5 and emphasized 9:6, understanding it to mean, contrary to the usual interpretation, "after an image, God made man," that is, in the Platonic sense of a heavenly archetype. He did not wish to allow any resemblance between God and any created being. Other interpretations sought to avoid the difficulty by rendering *elohim* (a plural form) not as "God" but as "divine beings" (*i.e.*, angels: "God created man after the image of divine beings [*elohim*]").

The earthly–spiritual creature. In those parts of the Jewish community of antiquity that were deeply influenced by Greek philosophical ideas, a dualistic interpretation of man was offered. Here the divine likeness suggested is that of the immortal, intellectual soul as contrasted to the body. Still other thinkers, both ancient and modern, have understood that likeness to be ethical, with particular emphasis placed on freedom of the will. What is evident is that no doctrine of man can be erected on the basis of these several verses alone, but that a broader view must be taken, in which they are assimilated.

"Soul," "spirit," or "life"

A careful examination of the biblical material, particularly the words *nefesh, neshama,* and *ruaḥ,* which are often too broadly translated as "soul" and "spirit," indicates that these must not be understood as referring to the psychical side of a psychophysical pair. A man did not possess a *nefesh* but rather was a *nefesh,* as Gen. 2:7 says: "*wayehi ha-adam le-nefesh ḥayya*" (". . . and the man became a living being"). Man was, for most of the biblical writers, what has been called "a unit of vital power," not a dual creature separable into two distinct parts of unequal importance and value. While this understanding of the nature of man dominated biblical thought, in apocalyptic literature (2nd century BCE–2nd century CE) the term *nefesh* began to be viewed as a separable psychical entity with existence apart from body. Although this was not entirely divorced from the unitary biblical view, nonetheless a functional body–soul dualism was present in such literature. In the Alexandrian version of Hellenistic Judaism the orientation toward Greek philosophy, particularly the Platonic view of the soul imprisoned in the flesh, led to a clear-cut dualism with a negative attitude toward the body. Rabbinic thought remained closer to the biblical position, at least in its understanding of man as a psychosomatic unit, although the temporary separation of the components after death was an accepted position.

The biblical view of man as an inseparable psychosomatic unit meant that death was understood to be his dissolution. Yet, although man ceased to be, this dissolution was not utter extinction. Some of the power that functioned in the unit may have continued to exist, but it was not to be understood any longer as life. The existence of the dead in *sheol,* the netherworld, was not living but the shadow or echo of living. For most of the biblical writers this existence was without experience, either of God or of anything else; it was unrelated to events. To call it immortality is to empty that term of any vital significance.

Sheol, resurrection, and immortality

However, this concept of *sheol,* along with belief in the possibility of occasional miraculous restorations of dead individuals to life, and perhaps even the idea of the revival of the people of Israel from the "death" of exile, provided a foothold for the development of belief in the resurrection of the dead body at some time in the future. The stimulus for this may have come from ancient Iranian religion, in which the dualistic cosmic struggle is eventually won by life through the resurrection of the dead. This idea began to appear in sketchy form in postexilic writings (Isa. 26:19; Dan. 12:2). In this view there is life only in the psychosomatic unit now restored. This restoration was bound up with the eschatological hope of Israel (see *Eschatology,* below) and was limited to the righteous. In subsequent apocalyptic literature a sharper distinction between body and soul was entertained, and the latter was conceived of as existing separately in a disembodied state after death. Although at this point the doctrine of the resurrection of the body was not put aside, nonetheless, the direction of thinking changed. The shades of *sheol* were now thought of as souls, and real personal survival—with continuity between life on earth and in *sheol*—was posited. Now Greek ideas, with their individualistic bent, began to have influence, so that the idea of resurrection that was in some way related to a final historical consummation, began to recede. True life after death was now seen as release from the bondage of the body, so that in place of, or alongside of, the afterlife of physical resurrection was set the afterlife of the immortal soul.

It was not the status of the soul, however, that concerned either the biblical or the rabbinic thinkers. What emerged from the latter's discussions of the biblical themes was an emphasis on the ethical import of man's composite makeup. Man is in a state of tension or equilibrium between the two foci of creation, the "heavenly" and the "earthly." He necessarily participates in both, and, as such, is the one responsible creature who can truly serve his Creator, for he alone, having both sides of creation in him, may choose between them. It is the ability to make an ethical choice that is the distinguishing mark of man. This ability is not derived from the "heavenly" side but resides in the double basis of man's existence. It is important to recognize this as something other than a body–soul dualism in which the soul is the source of good and the body the basis of evil. Such an attitude, however, did appear in some rabbinic material and was often affirmed in medieval philosophical and mystical speculations and by some of the later moralists. These are genuine variations and developments of the biblical-rabbinic ideas and may not be dismissed as aberrations. They represent authentic attempts to come to terms with other currents of thought and with the problems and uncertainties inherent within the earlier materials themselves.

The ethically bound creature. Mankind is then viewed, however this position is arrived at, as ethically involved. The first 11 chapters of Genesis are posited upon this responsibility, for the implicit assumption of the prepatriarchal stories is man's ability to choose between obedience and disobedience. Rabbinic Judaism, taking up the covenant-making episode between God and Noah (Gen. 9:8–17), developed it as the basis of mankind's ethical obligation. All men, not merely Israel, were engaged in a covenant relationship with God, which was spelled out in explicit precepts—variously enumerated as six, seven, or even 10 and occasionally as many as 30—that reflect general humanitarian behaviour and are intended to assure the maintenance of the natural order by the establishment of a proper human society. The Covenant with Israel was meant to bring into being a community that would advance the development of this society through its own obedience and witness.

Man's nature, viewed ethically, was explained in rabbinic Judaism not only as a tension between the "heavenly" and "earthly" components but also as a tension between two "impulses." Here again, fragmentary and allusive biblical materials were developed into more comprehensive statements. The biblical word *yetzer* means "plan," that which is formed in man's mind. In the two occurrences of the word in Genesis (6:5; 8:21), the plan or formation of man's mind is described as *raʿ*, perhaps "evil" in the moral sense or maybe no more than "disorderly," "confused," "undisciplined." The other biblical occurrences do not have this modifier. Nonetheless, the Aramaic translations (Targumim) invariably denominated it as *bisha* ("wicked") wherever it occurred. Rabbinic literature created a technical term *yetzer ha-raʿ* ("the evil impulse") to denote the source within man of his disobedience, and, subsequently, the counterterm *yetzer ha-ṭov* ("the good impulse") to indicate man's obedience. These more clearly suggest the ethical quality of man's duality, while their opposition and conflict point to man's freedom and the ethical choices he makes. Indeed, it is primarily within the realm of the ethical that Judaism posits freedom, recognizing the bound, or determined, quality of much of his existence (*e.g.*, his natural environment or physiological makeup).

The good and evil impulses

It is this ethically free creature who stands within the covenant relationship and who may choose to be obedient or disobedient. Sin, then, is ultimately deliberate disobedience or rebellion against the divine sovereign. This is more easily observed in relation to Israel, for it is here that the central concern of Judaism is most evident and the subject discussed in greatest detail. It should be noted, however, that since, according to Judaic tradition, all mankind stands within a covenant relation to God and is commanded to be moral and just, essentially the same choice is made universally. In technical language, the acceptance of divine sovereignty by the people of Israel and by the individual within that community is called "receiving the yoke of the kingship." This involves intellectual commitment to a basic belief, as expressed by the Deuteronomic proclamation: "Hear, O Israel, the Lord, our God, the Lord is one!" At the same time it imposes obligations in terms of communal and individual behaviour. These two responses are understood to be inextricably bound together, so that rejection of the divine sovereign is manifest as denial of God both intellectually and practically. It amounts to "breaking the yoke of the kingship." In more specific terms, sin is sometimes summed up under three major, interrelated headings: idolatry, murder, and illicit sexual behaviour, each and all of which explicitly and implicitly involve rebellion, for they involve activities that deny—if not God's existence—his commanding relationship and the requirement of man's response. Such behaviour destroys the community and sets individual against individual, thus thwarting the ultimate purpose of God, the perfected human society.

Teshuva ("turning")

If, however, man is free to choose rebellion and to suffer its consequences, he is also able to turn back to God and to become reconciled with him. The Bible—most particularly the prophetic writings—is filled with this idea, although the term *teshuva* ("turning") came into use only in rabbinic sources. Basically it grows out of the covenant and God's unwillingness—despite man's failures—to break off his relationship. In rabbinic thought it is apparently assumed that even the direst warnings of utter disaster and rejection imply the possibility of turning back to God, motivated by remorse and the desire for restoration. Divine readiness and human openness are the two sides of the process of reconciliation. What was expressed in prophetic literature in the immediate historicopolitical situation was stated in the synagogal liturgy in connection with pentateuchal and prophetic lessons and the homilies developed from them. Thus, the divine invitation was constantly being offered. Man was called upon to atone for his rebellion by positive action that repudiated his failure. He was summoned to reconstitute wholeness in his individual life and community in society.

Historically viewed, Jewish existence, as it developed under rabbinic leadership, following after the two disastrous rebellions against Rome, was an attempt to reconstitute a community of faith expressed in worship and in an ordered society that would enable the individual to live a hallowed life of response to the divine will. Although this plan was not spelled out in detail, it was probably understood to be the paradigm for the eventual reconstruction of humanity.

Medieval and modern views of man. The Jewish view of man is certainly less clearly articulated than its affirmations concerning God. Nonetheless, it is evident that its central concern was ethical. The question of how man as individual and community was to behave was the focus of interest. Yet it is also clear that metaphysical concerns, however rudimentary in the beginning, were included in the developing discussion. Medieval philosophers sought an accommodation between the doctrine of the resurrection of the body and the concept of the immortality of the soul. The greatest of them, Moses Maimonides (1135–1204), propounded an extremely subtle position that equated immortality with the cleaving of the human intellect to the active intellect of the universe, thus limiting it to philosophic adepts. In the modern period, the impact of various philosophical and psychological schools has further fragmented the situation so that little or no consensus is evident, although resurrection or immortality language is still used even when its content is uncertain. But alongside this lack of agreement, the view that man is to be understood, however else, as a creature who makes free ethical choices for which he is responsible remains—although variously articulated—the basic affirmation of Judaism about man.

ETHICS AND SOCIETY

The ethical emphasis of Judaism. Jewish affirmations about God and man intersect in the concept of Torah as the ordering of human existence in the direction of the divine. Man, however else understood, is an ethically responsible creature responsive to the presence of God in nature and in history. Although that responsiveness is expressed on many levels, it is within the horizontal relationship of man to man that it is most explicitly called for. The pentateuchal legislation sets down, albeit within the limitations of the structures of the ancient Middle East, the patterns of interpersonal relations. The prophetic messages are deeply concerned with these demands and see the disregard of them as the source of social and individual disorder. No segment of society, even the most exalted, is free of ethical obligation. Indeed, the transformation of prophetism from its earlier form as ecstaticism and soothsaying is seen in the ethical confrontation of David by Nathan ("Thou art the man") for seducing Bathsheba and arranging to have her husband killed (II Sam. 12). What is particularly striking is the affirmation that God is not only the source of ethical obligation but is himself the paradigm of it. In the so-called Code of Holiness (Lev. 19), it is imitation of divine holiness that is offered as the basis of human behaviour in the ethical sphere as well as the cultic-ceremonial. Concern for the economically vulnerable members of the community; obligations toward neighbours, hired labourers, and the physically handicapped; interfamilial relationships; and attitudes toward strangers (*i.e.*, non-Israelites) were all motivated by the basic injunction, "You shall be holy, for I, the Lord your God, am Holy." Acceptable human behaviour is, therefore, "walking in all His ways" (Deut. 11:22). The dialectic relation between God and man in the literary prophets also exhibits divine righteousness and divine compassion as patterns to be emulated in the life of the community.

The imitation of God

This theme, *imitatio Dei* ("imitation of God"), as developed in rabbinic Judaism, is expressed succinctly in a comment on the verse from Deuteronomy quoted above. In response to the question of how it is possible to walk "in all His ways," the reply is made (*Sifre* Deut. 85a): "As He is merciful and gracious, so be you merciful and gracious. As He is righteous so be you righteous. As He is holy, strive to be holy." Indeed even more daringly, God is described as clothing the naked, nursing the sick, comforting the mourners, burying the dead, so that man may recognize his own obligations.

Interpenetration of communal and individual ethics. What stands out in the entire development of Jewish ethical formulations is the constant interpenetration of communal and individual obligations and concerns. Although in the Book of Ezekiel (see especially chapter 18) emphasis is laid on individual responsibility, "the person who sins shall die," in contrast to the more widespread statement of communal involvement, "visiting the sins of the fathers upon the children" (*e.g.*, Ex. 20:5), these two aspects of ethical conduct are never entirely distinguished in Judaism. The just society requires the just man, and the just man functions within the just society. The concrete expression of ethical requirements in legal precepts took place with both ends in view, so that the process of beginning the holy community and the formation of the *ḥasid* ("pious"), the man of steadfast devotion to God, were concomitant processes. The relationship between the two is, of course, often mediated by the historical situation, so that in some periods one or the other moves to the centre of practical interest. In particular, the end of the Judaean state (70–135 CE) truncated the communal aspect of ethical obligations, often limiting discussion to apolitical responsibilities rather than to the full range of social involvements. The reestablishment of the State of Israel in the 20th century has, therefore, reopened for discussion areas that have for millennia been either ignored or relegated to the realm of abstraction. What this implies is that the full ethical responsibility of the Jew cannot be carried out solely within the realm of individual relationships but must include involvement in the life of a fully articulated community.

The just man in a just society

This double involvement is most vividly apparent in the biblical period, when both were equally present as divine command and demand. In the rabbinic period, because of the new political context, the communal aspect receded, so that discussion was mainly oriented toward the relationships between the members of the Jewish community or between individuals as such, and away from political responsibilities in the larger society. Nonetheless, the virtues that were understood to govern these relationships were, in their biblical setting, communal as well. Righteousness and compassion had been obligations of the state, governing the relationship between political units, as the first two chapters of Amos make evident. At the same time, as Micah 6:8 shows, doing justly, loving mercy, and walking humbly with God made up the pattern of the individual's obligations as well. Given the situation of the dispersion of the Jews following the revolts against Rome in the 1st and 2nd centuries CE, the individual pattern became the object of primary considerations. It is important to recognize that while theoretical ethical systems were not developed until the Middle Ages under the influence of philosophical concerns, nonetheless, even in the early period it was understood that behind the practical system of Halakha, the enumeration of legal precepts, there stood the dynamic of ethical theory. An attempt was made to reduce the hundreds of precepts to a small number expressing the ethical essence of Torah.

The key moral virtues. In keeping with the rabbinic understanding of Torah, study also was viewed as an ethical virtue. A passage in the traditional Prayer Book enumerates a series of virtuous acts—honouring parents, deeds of steadfast love, attendance twice daily at worship, hospitality to wayfarers, visiting the sick, dowering brides, accompanying the dead to the grave, devotion in prayer, peacemaking in the community and in family-life—and concludes by setting study of Torah as the premier virtue. Here is exhibited the complex variety of ethical behaviour called for within the Jewish tradition. To parental respect and family tranquillity are added, in other contexts, the responsibility of parents for children, the duties of husband and wife in the establishment and maintenance of a family, and ethical obligations that extend from the conjugal rights of each to the protection of the wife if the marriage is dissolved. The biblical description of God as upholding the cause of the fatherless and the widow and befriending the stranger, providing him with food and clothing (Deut. 10:18), remained a motivating factor in the structure of the community. Ethical requirements in economic life are expressed concretely in such a passage as Lev. 19:35–36: "You shall do no wrong in judgment, in measures of length or weight or quantity. You shall have just balances, just weights, a just *ephah*, and a just *hin*"; and in Amos' bitter condemnation of those who "sell the righteous for silver, and the needy for a pair of shoes" (Amos 2:6). Such injunctions, together with many other specific precepts and expressions of moral requirements, established the basis for a wide-ranging program that sought to govern, both in detail and in general, the economic life of the individual and the community. Not only are relations within the human sphere the object of ethical concern but nature also is so regarded. The animal world, in the biblical view, requires merciful consideration, so that not only man is commanded to rest on the Sabbath but his domestic animals are to share the rest with him (Ex. 20:10; 23:12). Mistreatment of beasts of burden is prohibited (Deut. 22:4); and wanton destruction of animal life falls under the ban (*ibidem:* 6–7). In the rabbinic attitude toward brute creation, even inanimate nature is the object of human solicitude. Thus, for example, the food-yielding trees of a city under siege may not be destroyed, according to Deuteronomic legislation (Deut. 20:14–20). The enlargement of this and other biblical precepts resulted in the generalized rabbinic prohibition "You shall not destroy" that governs man's use of his environment.

Protection of the weak

The relation to non-Jewish communities and cultures. As noted above, the end of the Jewish state reduced the scope of ethical judgments in the political sphere; nonetheless, relations between the Jewish community and other societies—particularly political units: the Roman and Christian empires, the Islāmic states, and other regimes—provided opportunities for the exploration of the ethical implications of such encounters. Since most of these were victor-victim, superior-inferior, power-powerless situations, with the Jews the weaker party, prudential considerations were dominant. Despite this, Jewish authorities sought to bring to bear upon these external arrangements the ethical standards that governed the internal structures.

The whole problem of the relationship between the Jewish community, in whatever form it has existed and does exist, and other social units has been vastly complicated. Ideally, the relation is that of witness to the divine intent in the world. Practically, it has swung between the extremes of isolation and assimilation, in which the ideal has, on occasion, been lost sight of. Culturally, from its earliest beginnings, the people Israel has met and engaged the ideas, forms, behaviour and attitudes of its neighbours in constructive development. It borrowed as it contributed and reformulated what it received in terms of its own commitments and affirmations. On more than a few occasions, as in the period of settlement in Canaan, it rejected the religiocultural ideas and forms of the native population. On others, it actively sought out—as in the Islāmic period in Spain (8th to 15th centuries)—ideas and cultural patterns of its neighbours, viewing them from its own perspective and embracing them when they were found to be of value. Indeed, the whole history of Israel's relationship with the world may be comprehended in the metaphor, used previously, of the heartbeat with its systole and diastole. No period of its existence discloses either total rejection of or abject surrender to other cultural and political structures but rather a tension, with the focal point always in motion at varying rates. Being more than a "confession" in the Christian sense, Judaism's adjustment to and relation with other sociopolitical units involved larger aspects of communal and individual life than merely the religious. Whether or not, under such circumstances, it is helpful to describe Judaism as a civilization, it is important to recognize that, viewed functionally, much more must be included than is usually subsumed under the common usage of "religion."

Isolation and assimilation

The formulation of Jewish ethical doctrines. The ethical concerns of Judaism have found frequent literary expression. Not only were rabbinic writings constantly directed toward the establishment of legal patterns that embody such concerns but in the medieval period the issues were dealt with in treatises on morals; in ethical wills, in which a father instructed his children about their obligations and behaviour; in sermons; and in other forms. In the 19th century the traditionalist Musar ("moral instructor") movement in eastern Europe and the philosophical discussions of the nascent Reform movement in the West focussed upon ethics. Indeed, since the political and social emancipation of the Jews, ethical and social rather than theological questions have tended to be given priority. Often the positions espoused have turned out to be, nonetheless, "judaized" versions of philosophic ethics or of political programs. In some instances, as in the case of the distinguished German-Jewish philosopher Hermann Cohen, the result has been a Jewishly compelling restatement of a secular philosophic ethics. In others, it has resulted in no more than a pastiche. More crucial, however, is the question of a unique Jewish ethics and of its authority. The reestablishment of a Jewish state renews the possibility that the full range of ethical decisions, including communal as well as individual responsibility, may be confronted. In such a situation, the ideal task of the people moves out of the realm of speculation to become actual again.

The Musar movement

THE UNIVERSE

Creation and Providence: God's world. Although the first chapter of Genesis affirms divine creation, it does not offer an entirely unambiguous view of the origin of the universe, as the debate over the correct understanding of Gen. 1:1 in former as in modern times discloses. (Was there or was there not a preexisting matter, void, or chaos?) Yet, basically, the interest of the author was not in the mode of creation, a later concern perhaps reflected in the various translations of the verse: "In the beginning God created," which could signify what medieval philosophers designated *creatio ex nihilo* ("creation out of nothing"); and "when God began to create," which could indicate some concept of prime matter. He was concerned rather to affirm that the totality of existence, inanimate (Gen. 1:3–19), living (20–25), and human (26–31), derived immediately from the same divine source; and, thus, that it is a universe. As divine creation, it is transparent to the presence of God, so that the Psalmist said: "The heavens declare the glory of God, and the expanse proclaims [that it is] the work of his hands" (19:1). Indeed, the repeated phrase: "And God saw how good it was" (Gen. 1:4, 10, 12, 18, 25, 31) may be understood as the ground of this affirmation, for the workmanship discloses the workman. The observed order of the universe is further understood by the biblical author to be the direct result of a covenantal relationship established between the world and God: "So long as the earth endures, seedtime and harvest. Cold and heat, summer and winter, day and night, shall not cease." (Gen. 8:22). This doctrine of the providential ordering of the universe, reaffirmed in rabbinic Judaism, is not without its difficulties, as in the liturgical change made in Isa. 45:7 to avoid ascribing evil to God. Nonetheless, despite the problem of theodicy (the problem of evil in a world made and ordered by God), Judaism has not acquiesced to the mood reported in the Palestinian Targum to Gen. 4:8: "He did not create the world in mercy nor does he rule in mercy." Rather, it has affirmed a benevolent and compassionate God.

Creation and history

It is the physical world—divine creation—that provides the stage for history, which is the place of the divine human encounter. An early Midrash, in response to the question as to why Scripture begins with the story of creation, points out that it was necessary in order to establish the identity of the Creator with the Giver of Torah, an argument basic to the liturgical structure of the Shema. This relationship is further emphasized in the Qiddush, the prayer of sanctification recited at the beginning of the Sabbath. That day is designated "a remembrance of creation" and "a recollection of the going-forth from Egypt." Thus, creation (nature) and history are understood to be inextricably bound up, for both derive from the same divine source. This being so, redemption—the reconciliation of God and man through and in history—does not ignore or exclude the natural world. Using the imagery of an extravagantly fecund world of nature, rabbinic thought expressed its view of the all-inclusive effects of the restored relationship.

Man's place in the universe. Man as creature is, of course, subject to the natural order. It is, indeed, in the world and through the world that man carries out his relationship to God. The commandments of Torah are obeyed not solely as observances between man and God but as actions between man and man, between man and the world. Although the creation story designates man as ruler over the earth and its inhabitants (Gen. 1:26–28; see also Ps. 8:5–9), nonetheless, far from being an arbitrary master, man's dominion is limited by Torah, for its regulations are concerned not only with transactions between man and man but also lay out his responsibilities to the land he cultivates, the produce of the soil, the animals he domesticates. Bound in the network of existence he, as the moral creature, is responsible for it in all of its parts.

Man's responsibility to nature

Even the destruction of the Jewish commonwealth in the 1st and 2nd centuries CE did not alienate the Jew from these responsibilities, as the elaborate system of Mishna and Gemara gives evidence. The gradual but consistent exclusion of the Jewish community from immediate connection with large segments of the natural world, through legislation in Christendom and Islām, tended to dull the Jew's awareness of it; the recurring references to it in the religious calendar, however, and the observation of harvest festivals even by citydwellers continued to remind the community of its ties. Thus, at the end of the 19th century, the nascent Zionist movement recognized that the regeneration of the Jewish people involved, among other requirements, a responsible relation to the natural order expressed in its attitude toward and treatment of the land.

As indicated in other contexts, the particular emphasis placed on one or the other side of the frequent twofoldness of the Jewish view has depended upon the situation in which the community has found itself. If nature as the place of divine disclosure has, during long periods of Jewish existence, assumed a somewhat subordinate role, it has never been rejected or been seen to be irrelevant to the divine purpose. Indeed, in Jewish eschatology, its restoration is part of the goal of history.

Intermediary beings: angels and demons. The exact nature of the nonhuman beings mentioned in Scripture—angels or messengers—is not altogether clear and their roles seem ephemeral. In the postexilic period, perhaps under Iranian influence, and in the late biblical and postbiblical literature, these beings emerge as more complete and often as clearly identifiable individuals with their own personal names. The unfocussed biblical view gave way to an elaborate hierarchy of functionaries who acted, in some apocalyptic visions, as a veritable heavenly bureaucracy. Nevertheless, despite a consensus concerning their existence, there was little agreement as to their role or importance. In some Midrashim God takes counsel with them; in other sources the rabbis urge men not to involve them but to approach God directly. Actually, they belong to that marginal area between religion and folklore. Like their counterfigures, the demons, they have a residual existence rooted in various layers of the Jewish experience and interpretation of the universe. At some times they are highly individualized and sharply realized; at others, they flit in and out of the imagination like bats in the evening. The medieval philosophers Aristotelized or Platonized them; the early mystics Neoplatonized them; the Kabbalists continually invented new ones and fitted them into their complicated network of cosmic existence. Nonetheless, their role, even in periods of considerable emphasis, was peripheral. They were outside the great movements and meanings of Jewish thought.

Contemporary philosophical speculation about the nature of the universe has, of course, required a response from Jewish thinkers. But, given the particular temper of a period in which metaphysics has not been central to much of theological discussion, no major statement has yet developed that has taken hold of the dominant positions and attempted to view them from the Jewish creationist perspective. The attempt within Reconstructionism to provide a naturalistic framework for Judaism, while courageous, lacks the breadth and depth of the great philosophical approaches.

ESCHATOLOGY

The future age of mankind and the world. The choice of Israel, according to the biblical writings, had occurred because of mankind's continual failure, by rebellion against its Creator, to fulfill its divine potential. The subsequent failure of Israel to become the holy community and thereby a witness to the nations gave rise to the prophetic movement that summoned the people to obedience. An integral part of prophetic summoning, side by side with threats of punishment and warnings of disaster, was the envisioning of a truly holy community, a society fully responding to the divine imperative. This kingdom of the future was conceived of as entirely natural, functioning as any normal sociopolitical unit and under the leadership of a human ruler, who would, however, carry out his tasks within the sphere of divine sovereignty, serving primarily to exhibit his own obedience and thus to stimulate the obedience of the entire people. This human monarch of the future was often, although not always, portrayed in terms of an idealized David, using such features of his life and reign as would underscore submission to God and emphasized social stability, economic satisfaction, and peace. During the period of the monarchy, the prophetic demand was directed toward each succeeding king, with the hope or even the expectation that he would be or become the new David, or the ideal ruler.

The Davidic model

The Babylonian Exile added a new measure of urgency to this expectation, although it was not expressed in any uniform fashion. The later chapters of the Book of Ezekiel provide in largely impersonal fashion the constitution for the new commonwealth but do not describe the peculiar characteristics of the ruler, while the later chapters of the Book of Isaiah focus on several figures—including Cyrus the Mede—who are seen as the divine instruments ushering in a new era. It is important to recognize that while such figures have extraordinary virtues ascribed to them, these virtues are neither superhuman nor suprahuman but such as are ultimately required of all Israel and of all men. The frustrations of the postexilic period, when the several attempts to bring into being the holy community had no more than partial success and were thwarted by the imperial designs of the great powers—as they had been in the preexilic period as well—led to an emphasis upon the futuristic quality of the messianic hope. This was abetted undoubtedly by external influences, such as Iranian thought, in which the cosmic rather than the historic aspect of a future era dominated. Since ancient cosmic myths—in good measure demythologized—had been part of the Israelite intellectual inheritance, evidenced at least in literary usages throughout Scriptures, the impact of such neighbouring ideas was to reinvigorate the mythic elements. Thus, hopes for the future at the end of the Persian period and on through the Hellenistic developments after *c.* 330 BCE comprised both historical expectations focussed upon a sociopolitical community and cosmic-mythic visions that moved on a broader stage. The latter were, of course, never entirely absent from the historical expectations and situations, for a renewal of nature was viewed as integral to the functioning of true society. The obedient community required, and was to be granted, a natural world in which true human relations could exist. In its most vivid forms, apocalypses (*i.e.*, visionary disclosures of the future), the literature of the period affords a remarkable insight into the agonies and urgencies of the people. After the failures, saving events, and disappointments of the past are recounted, the present, in transparent disguise, is portrayed and the immediately hoped-for intervention of God is described in awesome detail as a means of affirming and confirming the faith of those who saw themselves as the remnant, or perhaps the promise, of the holy community.

The socio-political and cosmic-mythic aspects

The king-messiah and his reign. Put schematically, Israel's hope was for the restoration of divine sovereignty over all of creation. Concretely, that hope found a considerable variety of expressions. Of all such expressions, that which centred around the idealized king began to assume an ever more important (but never exclusive) role. Many of the writings that report the ideas and attitudes of the Jewish community in the period immediately preceding and following the rise of Christianity are either ignorant of or more probably indifferent to the personal element. God is envisioned as the protagonist of the end, actively intervening or sending his messengers (*i.e.*, angels), to perform specific acts in ending the old and inaugurating the new era. On the other hand, in some writings of the period the anointed king-messiah (Hebrew, *mashiaḥ*, "anointed")—the title reflects the episode in I Sam. 16 in which David is thus singled out as the divinely chosen ruler—becomes more sharply defined as the central figure in the culminating events and, given the cosmic-mythic components, assumes suprahuman and in some instances, even quasi-divine, aspects. It is clear, then, that the doctrine of last things in Judaism is not necessarily messianic, if that term is properly limited to an inauguration of a future era through the action of a human, suprahuman, or quasi-divine person. Nonetheless, it must be recognized that the messianic version of eschatology played a more compelling role in rabbinic Judaism than other modes. The same is true with regard to the locus of the "world (or age) to come." Given the ingredients noted above, it was possible to construct various eschatological landscapes ranging from the mundane to the celestial, from Jerusalem in the hills of Judah to a heavenly city. Indeed, confronted with an embarrassment of riches, the medieval theologians sought to combine them into an inclusive system that intricately involved as large a variety of the possibilities as could be brought together. In such patterns the messianic this-worldly emphasis was understood as a preliminary movement toward an ultimate resolution. The ideal ruler, the new David, would reestablish the kingdom in its own land (in "Zion," or Palestine) and would reign in righteousness, equity, justice, and truth, thus bringing into being the holy nation and summoning all mankind to dwell under divine sovereignty. As a component of this reestablished kingdom, the righteous dead of Israel would be resurrected to enjoy life in the true community that did not exist in their days. This kingdom, however long it was destined to endure, was not permanent. It would come to an end either at a predetermined time or as victim of the unrepentant nations and cosmic foes, at which point the ultimate intervention by God would take place. All the wicked throughout history would be recalled to life, judged, and doomed; all the righteous would be transformed and transported into a new world; *i.e.*, creation would be totally restored. Particular emphases that one or the other of these ideas received, the ways in which they were interpreted—philosophically, mystically, or ethically—were determined most frequently by the situations and conditions in which the Jewish community found itself. With such a considerable body of ideas at its disposal and with the details of none of them ever receiving the kind of affirmation that statements about God, Torah, and Israel had, freedom of speculation in the realm of eschatology was little restricted. Thus, Joseph Albo (15th century) in his work on Jewish "dogmas," the *Sefer ha-ʿiqqarim*, was not inhibited from denying that belief in the messiah was fundamental. The mystical movements of the Middle Ages found in eschatological hopes a crucial centre. The early Kabbala was little interested in messianism, for it interiorized the expectations in the direction of personal redemption. Following the disasters of the late 15th to 17th centuries (*e.g.*, the expulsion of the Jews from Spain and the Cossack massacre of the Jews in Poland) however, messianic speculation in all of its varieties underwent a luxuriant growth, finally running wild in the movements surrounding Shabbetai Tzevi of Smyrna and later Jacob Frank of Offenbach. These tragedies for the Jewish communities once again resulted in a futurizing of the hopes or at least a limiting of their application (see also below, *Jewish mysticism*).

Resurrection of the dead and Last Judgment

Secularization of messianism. In the 19th century, with the political emancipation of the Jews in western Europe and the development of an optimistic evolutionism, messianism was transformed by many liberal thinkers into a version of the idea of progress whose goal was often thought of as immediately attainable through enlightened social and political action. When disillusionment with the emancipation set in, messianism was even more completely secularized in some segments of the community who saw its meaning and fulfillment in some form of socialism—again, rather close at hand. In others, it was absorbed into the emerging political nationalism—Zionism. Similar developments took place in eastern Europe, with parallel transformations. In more recent times, particularly since the events symbolized by the name Auschwitz (a Nazi death camp in Poland, where millions of Jews were exterminated), the earlier modern interpretations, particularly of messianism, but also of eschatology as a whole, have been considered inadequate. Although no compelling statement has been forthcoming, Jewish thinkers in the second half of the 20th century have been attempting once again to come to grips with eschatological concepts in all of their varieties and forms.

Basic practices and institutions

THE HALLOWING OF EVERYDAY EXISTENCE

Systematic presentations of the affirmations of the Jewish community never served as the sole mode of expressing beliefs of the people. Side by side with speculation—Haggadic, philosophic, mystical, or ethical—there stood, not in a secondary role but as the other of the double focuses, Halakha ("practice," "rules of conduct"), the paradigmatic statement of the behaviour, individual and communal, that embodied concretely the beliefs conceptualized in speculation. Life in the holy community was understood to embrace every level of human existence. The prophets vigorously resisted attempts to limit the sovereignty of the God of Israel to organized worship and ritual. The Pharisees, even while the Jerusalem Temple cult was still in existence, sought to reduce priestly exclusiveness by enlarging the scope of sacral rules to include, as far as possible, all of the people. Rabbinic Judaism, Pharisaism's surviving descendant, continued the process of democratization and sought, through its system of interpretation, to find in every occasion of life a means of affirming divine concern and presence. Viewed negatively by some Protestant theologians, this development has been judged to stifle spontaneity. Yet spontaneity is not necessarily lacking in a world governed by Halakha, although the danger of the stylized routine in religious and ethical life is apparent. Nonetheless, the intention of the Halakhic attitude is to remind the Jew constantly that each and every occasion of life is a locus of divine disclosure. This is most clearly seen in the *berakhot*, the "blessings," that are prescribed to accompany the performance of a broad spectrum of human actions, from the commonplace routines of daily life to the restricted gestures of the cultic-liturgical year In these, God is addressed directly in the second person singular, his sovereignty is affirmed, and his activity as Creator, Giver of Torah, or redeemer, expressed in a wide variety of eulogies, is proclaimed. There are no areas of human behaviour in which man cannot be met by God, and in terms of its intention, the Halakhic pattern is designed to make such possibilities experienced realities. Yet, again, it must be noted that the situation of the Jewish community determines in a very large way how the intention is actualized. On more than one occasion the Halakhic pattern has served as a defense against a hostile environment and has thus tended to become scrupulosity (an obsessive concern with minute details), but the dynamic of the intention itself has as often broken through to re-establish its integrity and hallow life in its wholeness.

The centrality of Halakha

THE TRADITIONAL PATTERN OF INDIVIDUAL AND FAMILIAL PRACTICES

Perspective on the traditional pattern of an individual's life is obtained by examining a passage from the Babylonian Talmud (tractate *Berakhot* 60b) that was subsequently reworked into a liturgical structure but which in its original form exhibits the intention discussed above. In this passage, the blessings accompanying a man's waking and returning to the routines of life are prescribed. There is a brief thanksgiving on awakening for being restored to conscious life; then the impingement of the external world is responded to in a benediction over the cock's crowing; following this, each ordinary act, opening one's eyes, stretching and sitting up, dressing, standing up, walking, tying one's shoes, fastening one's belt, covering the head, washing the hands and face, has its accompanying blessing, reminding a man that the world and the life to which he has returned exist in the presence of God. These are followed by a supplication in which the petitioner asks that his life during the day may be worthy in all of its relationships. Then, as the first order of daily business, Torah, both written (Bible) and oral (Mishna), is briefly studied, introduced by eulogies of God as Giver of Torah. Finally, there is a prayer for the establishment of the Kingdom of God, for each day contains within itself the possibility of ultimate fulfillment. As indicated, this was originally not a part of public worship (even today it is, strictly speaking, not part of the synagogue service, although it is most frequently recited there) but was personal preparation for a life to be lived in the presence of God.

The liturgy of daily life

From H. Hapgood, *The Spirit of the Ghetto*, © 1965 by Funk & Wagnalls Co., Inc., reproduced by permission of Funk & Wagnalls Publishing Co., Inc

Boy putting on phylacteries in preparation for morning prayer, drawing by Jacob Epstein, *c.* 1902.

Such individual responsibility marks much of Jewish observance, so that the synagogue—far from being the focus of observance—shares with the home and the workaday world the opportunities for the divine-human encounter. The table blessings, Qiddush (the "sanctification" of the Sabbath and festivals), the erection of the booth (*sukka*) for Sukkot (the Feast of Tabernacles), the seder (the festive Passover meal) with its symbols and narration of the Exodus, the lighting of the lamps during the eight days of Ḥanukka (the Feast of Dedication), are all the obligation of the individual and the family and have their place in the home. It is here, too, that woman's role is defined and here, as contrasted with the synagogue, that she functions centrally. Given the traditional dietary regimen of the Jewish community—the exclusion of swine, carrion eaters, shellfish, and other creatures, the separation of meat and dairy products, the ritual slaughtering of animals, the required separation and burning of a small portion of dough (*ḥalla*) when baking, the supervision of the Passover food requirements, and many other stipulations—there exists a large and meticulously governed

The central role of women in family religion area within the home that is indeed the sphere of woman's religion. There seems not to have been a hierarchy of values in which the home-centred, as contrasted with the synagogue-oriented, practices were given an inferior status. In modern times, however, particularly in Western civilization where the pervasiveness of religious obligation has been replaced by ecclesiastical institutionalism, on the prevailing Christian model, this whole crucial area has lost much of its meaning as a place of divine-human meeting. Thus, for many it is only the synagogue that provides such an opportunity, and the individual act has been reduced on the scale of values. With this downgrading, woman's religion has lost its significance so that her status—when parallels are drawn to her role in the larger society—has been reduced to one of inferiority. However attenuated personal religious responsibility may have become in some environments or transformed into stylized cultural forms, the intention that informs the Halakhic structure, the hallowing of the individual's total existence, remains a potent force within the Jewish community.

THE TRADITIONAL PATTERN OF SYNAGOGUE PRACTICES

The other focus of observance is the synagogue. The origins of this institution are obscure and a number of hypotheses have been proposed to account for the appearance of this essentially lay-oriented form of worship. What seems certain is that during the period of the Second Temple—following the return from Babylon and continuing until the Temple destruction in 70 CE—there were, side by side with the official cult, other modes of worship more or less independent of the priesthood and nonsacrificial in form. The reports by the philosopher Philo and the historian Josephus in the 1st century, buttressed by the Qumrān document (Dead Sea Scrolls), provide some knowledge of the practices of the contemporary Essenes; rabbinic sources, including the earliest layers of the traditional order of worship, enable us to understand another, apparently Pharisaic, mode; the brief allusions to the practices of James and his Jewish Christian companions in the book of Acts suggest yet other varieties. In any case, the grouping that formed the cadre of what eventually became rabbinic Judaism observed some form of worship that, with the destruction of the Temple cult, was able to provide a new centre and even to absorb enough from the defunct priestly institution to suggest continuity and legitimacy. This was probably the basic pattern for synagogal liturgy in the millennia that followed.

Readings from the Torah and prophets

At the heart of synagogal worship is the public reading of Scriptures. This takes place at the morning service on Sabbaths, holy days, and festivals, on Monday and Thursday mornings, and on Sabbath afternoons. The readings from the Pentateuch are presently arranged in an annual cycle so that, beginning on the Sabbath following the autumnal festivals with Gen. 1:1, the entire five books are read through the rest of the year. The texts for festivals, holy days, and fasts reflect the particular significance of those occasions. In addition, a second portion from the prophetic writings (in the Jewish tradition these include Joshua, Judges, Samuel, and Kings, as well as the three major and 12 minor prophets, but not Daniel) is read on many of these occasions. All of this takes place within the structure of public worship and is provided with ceremonies during which the Sefer Torah ("Book of the Torah"), the pentateuchal scroll, is removed from the ark (cabinet) at the front of the synagogue, and carried in procession to the reading desk; from it, the pertinent text is chanted by the reader. The text for the service is divided into subsections varying from seven on the Sabbath to three at the weekday morning service, and individuals are called forward to recite the blessings eulogizing God as Giver of Torah before and after each of these. The order of worship is composed of the preparatory blessings and prayers noted above, to which are added passages recalling the Temple sacrificial cult (thus relating the present form of worship to the past); the recitation of a number of Psalms and biblical prayers; the Shema and its accompanying benedictions, introduced by a call to worship that marks the beginning of formal public worship; the prayer (*tefilla*) in the strict sense of petition; confession and supplication (*taḥanun*) on weekdays; the reading of Scripture; and concluding acts of worship. This general structure of the morning service varies somewhat, with additions and subtractions for the afternoon and evening services and for Sabbath, holy days, and festivals.

The "Eighteen Benedictions"

The prayer (*tefilla*), just mentioned, is often called the *shemone 'esre*, the "Eighteen Benedictions"—although it actually has 19—or the *'amida*, "standing," because it is recited in that position. It is made up of three introductory benedictions: praise of the God of the Fathers, of God the Redeemer who resurrects the dead, and of God the holy one who fills the earth with his glory, and of three concluding acts; a prayer for the acceptance of the service, a thanksgiving, and a prayer for peace—with a series of intermediate petitions for knowledge, well-being, acceptance of repentance, forgiveness of sin, and others. On the Sabbath and festivals these are replaced by benedictions that mention the specific occasion but are not petitionary, it being considered inappropriate to attend to workday concerns at these times.

While the general outline of this order of service is found throughout the entire Jewish world, the details have varied, both in different periods and in geographic and cultural areas. The public service, requiring the presence of at least 10 males, the *minyan* ("quorum"), is generally led by a synagogal official, the *ḥazzan*, or cantor, but any Jewish male with the requisite knowledge may act in this capacity since there is, quite strictly, no clerical class in the community to whom such leadership is limited (see *The rabbi*, below).

The synagogue room itself has a very simple basic form although, of course, it may be embellished considerably. The only requirements are a container for the Torah scroll(s), the *aron ha-qodesh* ("the holy ark")—a chest against the east wall, or a recessed closet with doors and a curtain; a prayer desk (*'amud*) facing the ark at which the reader stands when reciting the service; and the pulpit (*bima*)—according to some requirements in or close to the centre of the room—from which the Torah is read. In the Spanish-Portuguese tradition, only one desk (called *teva*) is used. The ark contains one or more scrolls, on which are written the five books of Moses. These are variously ornamented, depending upon the cultural region: European communities decking them in coverings of cloth; Oriental (North African and Near Eastern) placing them in wooden or metal containers. In addition, silver ornaments, in the form of towers or crowns, are often set on the tops of two rods on which the scroll is wound, and a breastplate and a pointer are suspended from them.

Accommodations for the worshippers vary according to the cultural milieu, from rugs and cushions in Oriental synagogues to pews and standing desks in European ones. Given this essential simplicity, the synagogue room itself may be used for other purposes than worship, *e.g.*, study and community assembly. Again, this varies with the cultural pattern.

CEREMONIES MARKING THE INDIVIDUAL LIFE CYCLES

There are within Jewish life two cycles corresponding to the individual and the synagogal focuses, although they necessarily intermingle. The life of the individual is marked by observances that single out the notable events of personal existence. A male child is circumcised on the eighth day following birth, as a covenantal sign (Gen. 17); the rite of circumcision (*berit mila*) is accompanied by appropriate benedictions and ceremonies, including naming. Females are named in the synagogue, generally on the Sabbath following birth, when the father is called to recite the benedictions over the reading of Torah. A firstborn son, if he does not belong to a priestly or a levitical family, is redeemed at one month (in accordance with Ex. 13:12–13 and Num. 18:14–16) by the payment of a stipulated sum to a cohen (a putative member of the priestly family). On arrival at the age of 13, a boy is called publicly to recite the Torah benedictions, thus signifying his religious coming-of-age; he is thenceforth obligated to observe the commandments as his own responsibility—he is now a Bar Mitzwa ("Son of the Commandment"). Marriage (*ḥatuna*, also *Qiddushin*, "sanctifications") involves

Circumcision and naming

Boy reading the Torah at synagogue services, an important part of the Bar Mitzwa ceremony.
Cornell Capa—Magnum

a double ceremony, performed together in modern times but separated in ancient times by a year. First is the betrothal (*erusin*), which includes the reading of the marriage contract (*ketubba*) and the giving of the ring with a declaration, "Behold you are consecrated to me by this ring according to the law of Moses and Israel," accompanied by certain benedictions. This is followed by the marriage proper (*nissu'in*), consisting of the reciting of the seven marriage benedictions. The ceremony is performed under a *ḥuppa*, a canopy, that symbolizes the bridal bower.

The burial service is marked by simplicity. The body is prepared for the grave by the *ḥevra' qaddisha'* (the holy society), clad only in a simple shroud, and the interment takes place as soon after death as possible. In Israel no coffin is used. There are observances connected with death, many of which belong to the realm of folklore rather than Halakhic tradition. A mourning period of 30 days is observed, of which the first seven (*Shiv'a*) are the most rigorous. During the 11 months following a death, the bereaved recite a particular form of a synagogal doxology (Qaddish) during the public service as an act of memorial. The doxology itself, entirely devoid of any mention of death, is a praise of God and a prayer for the establishment of the coming Kingdom. It is also recited annually on the anniversary of the death (*yahrzeit*). (L.H.S.)

THE CYCLE OF THE RELIGIOUS YEAR

The term Jewish religious year as used in this section encompasses the cycle of Sabbaths and holidays that are commonly observed by the Jewish religious community—and officially in Israel by the Jewish secular community as well. The Sabbath and festivals are bound to the Jewish calendar, reoccur at fixed intervals, and are celebrated at home and in the synagogue according to ritual set forth in Jewish law and hallowed by Jewish custom. According to Jewish teaching, the Sabbath and festivals are, in the first instance, commemorative. The Sabbath, for example, commemorates the Creation, and Passover commemorates the Exodus from Egypt over 3,000 years ago. The past is not merely recalled; it is also relived through the Sabbath and festival observances. Creative physical activity ceases on the Sabbath as it did, according to Genesis, when the Creation was completed; Jews leave their homes and reside in booths during the Sukkot festival as did their biblical ancestors. Moreover, Sabbath and festival themes are considered to be perpetually significant, recurring and renewed in every generation. Thus the revelation of the Torah (the divine teaching or law) at Sinai, commemorated on Shavuot, is considered an ongoing process which recurs whenever a commitment is made to Torah study.

An important aspect of Sabbath and festival observance is sanctification. The Sabbath and festivals sanctified the Jews more than the Jews sanctified the Sabbath and festivals. Mundane meals became sacred meals; joy and relaxation became sacred obligations (*mitzwot*). No less significant is the contribution of the Sabbath and festivals toward communal awareness. Thus, neither Sabbath nor festival can be properly observed in the synagogue according to the ancient tradition if fewer than 10 male Jews are present. Again, a Jew prays on Rosh Hashana and mourns on Tisha be-Av not only for his own fate but for the fate of all Jews. The sense of social cohesiveness fostered by the Sabbath and festival observances has stood the Jews well throughout their long, often tortuous history.

The seven-day week, the notion of a weekly day of rest, and many Christian and Islāmic holiday observances owe their origins to the Jewish calendar, Sabbath, and festivals.

The Jewish calendar. *Lunisolar structure.* The Jewish calendar is lunisolar—*i.e.*, regulated by the positions of both the moon and the sun. It consists usually of 12 alternating lunar months of 29 and 30 days each (except for Ḥeshvan and Kislev, which sometimes have either 29 or 30 days), and totals 353, 354, or 355 days per year. The average lunar year (354 days) is adjusted to the solar year (365¼ days) by the periodic introduction of leap years in order to assure that the major festivals fall in their proper season. The leap year consists of an additional 30-day month called First Adar, which always precedes the month of (Second) Adar. A leap year consists of either 383, 384, or 385 days and occurs seven times during every 19-year period (the so-called Metonic cycle). Among the consequences of the lunisolar structure are these: (1) The number of days in a year may vary considerably, from 353 to 385 days. (2) The first day of a month can fall on any day of the week, that day varying from year to year. Consequently, the days of the week upon which an annual Jewish festival falls vary from year to year despite the festival's fixed position in the Jewish month.

Leap year: Second Adar

Months and notable days. The months of the Jewish religious year, their approximate equivalent in the Western Gregorian calendar, and their notable days, are as follows:

Tishri (September–October)
- 1, 2 Rosh Hashana (New Year)
- 3 Tzom Gedaliahu (Fast of Gedaliah)
- 10 Yom Kippur (Day of Atonement)
- 15–21 Sukkot (Tabernacles)
- 22 Shemini Atzeret (Eighth Day of the Solemn Assembly)
- 23 Simḥat Torah (Rejoicing of the Law)

Ḥeshvan, or Marḥeshvan (October–November)

Kislev (November–December)
- 25 Ḥanukka (Feast of Dedication) begins

Ṭevet (December–January)
- 2–3 Ḥanukka ends
- 10 'Asara be-Tevet (Fast of Tevet 10)

Shevaṭ (January–February)
- 15 Tu bi-Shevaṭ (15th of Shevaṭ: New Year for Trees)

Adar (February–March)
- 13 Ta'anit Esther (Fast of Esther)
- 14, 15 Purim (Feast of Lots)

Nisan (March–April)
- 15–22 Pesaḥ (Passover)

Iyyar (April–May)
- 18 Lag ba-Omer (33rd Day of the Omer Counting)

Sivan (May–June)
- 6, 7 Shavuot (Feast of Weeks, or Pentecost)

Tammuz (June–July)
- 17 Shiva' 'Asar be-Tammuz (Fast of Tammuz 17)

Av (July–August)
- 9 Tisha be-Av (Fast of Av 9)

Elul (August–September)

During leap year, the Adar holidays are postponed to Second Adar.

Since 1948 many Jewish calendars list Iyyar 5—Israel Independence Day—among the Jewish holidays.

Origin and development. The origin of the Jewish calendar can no longer be accurately traced. Some scholars suggest that a solar year prevailed in ancient Israel, but no convincing proofs have been offered, and it is more likely that a lunisolar calendar similar to that of ancient Babylonia prevailed in ancient Israel. In late Second Temple times (*i.e.*, 1st century BCE to 70 CE), calendrical matters were regulated by the Sanhedrin, or council of elders, at Jerusalem. The testimony of two witnesses who had observed the New Moon was ordinarily required to proclaim a new month. Leap years were proclaimed by a council of three or more rabbis with the approval of the *nasi*, or president, of the Sanhedrin. With the decline of the Sanhedrin, calendrical matters were decided by the Palestinian patriarchate (the official heads of the Jewish community under Roman rule). Jewish persecution under Constantius II (reigned 337–361) and advances in astronomical science led to the gradual replacement of observation by calculation. According to Hai ben Sherira (died 1038)—the head of a leading Talmudic academy in Babylonia—Hillel II, a Palestinian patriarch, introduced a fixed and continuous calendar in 359 CE. A summary of the regulations governing the present calendar is provided by Maimonides, the great medieval philosopher and legist, in his *Code: Sanctification of the New Moon*, chapters 6–10.

Regulation by the Sanhedrin and patriarchate

Fragments of writings discovered in a geniza (depository for sacred writings withdrawn from circulation) have brought to light a calendrical dispute between Aaron ben Meir, a 10th-century Palestinian descendant of the patriarchal (Hillel) family, and the Babylonian Jewish authorities, including Saʿadia ben Joseph—an eminent 10th-century philosopher and *gaon* (head of a talmudic academy). Ben Meir's calculations provided that Passover in 922 be celebrated two days earlier than the date fixed by the normative calendar. After a bitter exchange of letters, the controversy subsided in favour of the Babylonian authorities, whose hegemony in calendrical matters was never again challenged.

Calendars of various sectarian Jewish communities deviated considerably from the normative calendar described above. The Dead Sea (or Qumrān) community (made famous by the Dead Sea Scrolls discoveries) adopted the calendrical system of the noncanonical books of *Jubilees* and *Enoch*, which was essentially a solar calendar. Elements of this same calendar reappear among the Mishawites, a sect founded in the 9th century.

The Karaites, a sect founded in the 8th century, refused, with some exceptions, to recognize the normative fixed calendar and reintroduced observation of the New Moon. Leap years were determined by observing the maturation of the barley crop in Palestine. Consequently, Karaites often celebrated the festivals on dates different from those fixed by the rabbis. Later, in medieval times, the Karaites adopted some of the normative calendrical practices, while rejecting others.

The Sabbath. The Jewish Sabbath (from Hebrew *shavat*, "to rest") is observed throughout the year on the seventh day of the week—Saturday. According to biblical tradition, it commemorates the original seventh day on which God rested after completing the creation.

Scholars have not succeeded in tracing the origin of the seven-day week, nor can they account for the origin of the Sabbath. A seven-day week does not accord well with either a solar or lunar calendar. Some scholars, pointing to the Akkadian term *shapattu*, suggest a Babylonian origin for the seven-day week and the Sabbath. But *shapattu*, which refers to the day of the Full Moon and is nowhere described as a day of rest, has little in common with the Jewish Sabbath. It appears that the notion of the Sabbath as a holy day of rest, linking God to his people and recurring every seventh day, was unique to ancient Israel.

Importance. The central significance of the Sabbath for Judaism is reflected in the traditional commentative and interpretative literature called Talmud and Midrash (*e.g.*, "if you wish to destroy the Jewish people, abolish their Sabbath first") and in numerous legends and adages from more recent literature (*e.g.*, "more than Israel kept the Sabbath, the Sabbath kept Israel"). Some of the basic teachings of Judaism affirmed by the Sabbath are God's acts of creation, God's role in history, and God's covenant with Israel. Moreover, the Sabbath is the only Jewish holiday the observance of which is enjoined by the Ten Commandments. Jews are obligated to sanctify the Sabbath at home and in the synagogue by observing the Sabbath laws and engaging in worship and study. The leisure hours afforded by the ban against work on the Sabbath were put to good use by the rabbis, who used them to promote intellectual activity and spiritual regeneration among Jews. Other days of rest, such as the Christian Sunday and the Islāmic Friday, owe their origins to the Jewish Sabbath.

Worship, study, and abstinence from work

Observances. The biblical ban against work on the Sabbath, while never clearly defined, includes such activities as baking and cooking, travelling, kindling fire, gathering wood, buying and selling, and bearing burdens from one domain into another. The Talmudic rabbis listed 39 major categories of prohibited work, including agricultural activity (*e.g.*, plowing and reaping), work entailed in the manufacture of cloth (*e.g.*, spinning and weaving), work entailed in preparing documents (*e.g.*, writing), and other forms of constructive work.

At home, the Sabbath begins Friday evening some 20 minutes before sunset, with the kindling of the Sabbath candles by the wife, or in her absence by the husband. In the synagogue, the Sabbath is ushered in at sunset with the recital of selected psalms and the *Lekha Dodi*, a 16th-century Kabbalistic (mystical) poem. The refrain of the latter goes: "Come, my beloved, to meet the bride." the "bride" being the Sabbath. After the evening service, each Jewish household begins the first of three festive Sabbath meals by reciting the *Qiddush* ("sanctification" of the Sabbath) over a cup of wine. This is followed by a ritual washing of the hands and the breaking of bread; two loaves of bread (commemorating the double portions of manna described in Exodus) being placed before the breaker of bread at each Sabbath meal. After the festive meal, the remainder of the evening is devoted to study or relaxation. The distinctive features of the Sabbath morning synagogue service include the public reading of the Torah, or Five Books of Moses (the portion read varies from week to week) and, generally, the sermon, both of which serve to educate the listeners. Following the service, the second Sabbath meal begins, again preceded by *Qiddush* (of lesser significance), and conforming for the most part to the first Sabbath meal. The afternoon synagogue service is followed by the third festive meal (without *Qiddush*). After the evening service, the Sabbath comes to a close with the Havdala ("Distinction") ceremony, which consists of a benediction noting the distinction between Sabbath and weekday, usually recited over a cup of wine accompanied by a spice box and candle.

The Havdala ceremony

The Jewish holidays. The major Jewish holidays are the Pilgrim Festivals: Pesaḥ (Passover), Shavuot (Feast of Weeks, or Pentecost), and Sukkot (Tabernacles); and the High Holidays: Rosh Hashana (New Year) and Yom Kippur (Day of Atonement). In common, their observance is required by the Torah and work is prohibited for the duration of the holiday (except on the intermediary days of the Pesaḥ and Sukkot festivals, when work the neglect of which entails monetary loss is permitted). Purim (Feast of Lots) and Ḥanukka (Feast of Dedication), while not mentioned in the Torah (and therefore of lesser solemnity), were instituted by Jewish authorities in the Persian and Greco-Roman periods. Lacking the work restrictions characteristic of the major festivals, they are sometimes regarded as minor festivals. In addition, there are the five fasts: ʿAsara be-Ṭevet (Fast of 10 Ṭevet), Shivaʿ ʿAsar be-Tammuz (Fast of Tammuz 17), Tisha be-Av (Fast of Av 9), Tzom Gedaliahu (Fast of Gedaliah), and Taʿanit Esther (Fast of Esther); and the lesser holidays—*i.e.*, holidays the observances of which are few and not always clearly defined—such as Rosh Ḥodesh (First Day of the Month), Ṭu bi-Shevaṭ (New Year for Trees), and Lag ba-ʿOmer (33rd Day of Omer Counting). The fasts and the lesser holidays also lack the work restrictions characteristic of the major festivals. Some of the fasts and Rosh Hodesh are mentioned in Scripture, but most of the details concerning their proper observance, as well as those concerning the other lesser holidays, were provided by the Talmudic and medieval rabbis.

Pilgrim festivals. In Temple times, all males were required to appear at the Temple three times annually and actively participate in the festal offerings and celebrations. These were the joyous pilgrim festivals of Pesaḥ, Shavuot, and Sukkot. Originally, they marked the major agricultural seasons in ancient Israel and commemorated Israel's early history; but after the destruction of the Second Temple in 70 CE, emphasis was almost exclusively placed on the commemorative aspect.

In modern Israel, Pesaḥ, Shavuot, and Sukkot are celebrated for the number of days prescribed by Scripture, namely, seven days, one day, and eight days, respectively (with Shemini Atzeret added to Sukkot). Due to calendrical uncertainties which arose in Second Temple times (6th century BCE to 1st century CE), each festival is celebrated for an additional day in the Diaspora.

Pesaḥ (Passover)

Pesaḥ commemorates the Exodus from Egypt and the servitude that preceded it. As such, it is the most significant of the commemorative holidays, for it celebrates the very inception of the Jewish people—*i.e.*, the event which provided the basis for the covenant between God and Israel. The term *pesaḥ* refers originally to the paschal (Passover) lamb sacrificed on the eve of the Exodus, the blood of which marked the Jewish homes to be spared from God's plague; its etymological significance, however, remains uncertain. The Hebrew root is usually rendered "passed over"—*i.e.*, God passed over the homes of the Israelites when inflicting the last plague on the Egyptians—hence the term Passover. The festival is also called Ḥag, Matzot ("Festival of Unleav- ened Bread"), for unleavened bread is the only kind of bread consumed during Passover.

Leaven (*se'or*) and foods containing leaven (*ḥametz*) are neither to be owned nor consumed during Pesaḥ. Aside from meats, fresh fruits, and vegetables, it is customary to consume only those foods prepared under rabbinic supervision and labelled "kosher for Passover," warranting that they are completely free of contact with leaven. In many homes, special sets of crockery, cutlery, and cooking utensils are acquired for Passover use. On the evening preceding the 14th day of Nisan, the home is thoroughly searched for any trace of leaven (*bediqat ḥametz*). The following morning the remaining particles of leaven are destroyed by fire (*bi'ur ḥametz*). From then until after Pesaḥ, no leaven is consumed. Many Jews sell their more valuable leaven products to non-Jews before Passover (*mekhirat ḥametz*), repurchasing the foodstuffs immediately after the holiday.

The unleavened bread (*matza*) consists entirely of flour and water, great care being taken to prevent any fermentation before baking. Hand-baked *matza* is flat, rounded, and perforated. Since the 19th century, many Jews have preferred the square-shaped, machine-made *matza*.

Passover eve is ushered in at the synagogue service on the evening before Passover, after which each family partakes of the seder ("order of service); *i.e.*, an elaborate festival meal in which every ritual is regulated by the rabbis. (In the Diaspora, the seder is also celebrated on the second evening of Passover.) The table is bedecked with an assortment of foods symbolizing the passage from slavery (*e.g.*, bitter herbs) into freedom (*e.g.*, wine). The Haggada (literally "narration"), a printed manual comprised of appropriate passages culled from Scripture, Talmud, and Midrash, accompanied by medieval hymns, serves as a guide for the ensuing ceremonies and is recited as the evening proceeds. The seder opens with the cup of sanctification (Qiddush), the first of four cups of wine drunk by the celebrants. An invitation is extended to the needy to join the seder ceremonies, after which the youngest son asks four prescribed questions expressing his surprise at the many departures from usual mealtime procedure. ("How different this night is from all other nights!") The father then explains that the Jews were once slaves in Egypt, were then liberated by God, and now commemorate the servitude and freedom by means of the seder ceremonies. Special blessings are recited over the unleavened bread and the bitter herbs (*maror*), after which the main courses are served. The meal closes with a serving of *matza* recalling the paschal lamb, consumption of which concluded the meal in Temple times. The seder concludes with the joyous recital of hymns praising God's glorious acts in history and anticipating a messianic redemption to come.

The Passover liturgy is considerably expanded and includes the daily recitation of Psalms 113–118 (Hallel, "praise"), public readings from the Torah, and an additional service (*musaf*). On the first day of Pesaḥ, a prayer for dew in the Holy Land is recited; on the last day, the memorial service for the departed (*yizkor*) is added.

Shavuot (Feast of Weeks, or Pentecost)

Originally an agricultural festival marking the wheat harvest, Shavuot commemorates the revelation of the Torah at Sinai. Shavuot ("weeks") takes its name from the seven weeks of grain harvest separating Passover and Shavuot. The festival is also called Ḥag ha-Qazir (Harvest Festival) and Yom ha-Bikkurim (Day of First Fruits). Greek-speaking Jews called it *pentēkostē*, meaning "the fiftieth" day after the sheaf offering. In rabbinic literature, Shavuot is called *atzeret* ("cessation, conclusion"), perhaps because the cessation of work is one of its distinctive features, or possibly because it was viewed as concluding the Passover season. In liturgical texts it is described as the "season of the giving of our Torah." The association of Shavuot with the revelation at Sinai, while not attested to in Scripture, is alluded to in the Pseudepigrapha (a collection of noncanonical writings). In rabbinic literature the association first appears in 2nd-century materials. The association, probably an ancient one, was derived in part from the book of Exodus, which dates the revelation at Sinai to the third month (counting from Nisan), *i.e.*, Sivan.

Scripture does not provide an absolute date for Shavuot. Instead, 50 days (or seven weeks) are reckoned from the day the sheaf offering ('Omer) of the harvest was brought to the Temple, the 50th day being Shavuot. According to the Talmudic rabbis, the sheaf offering was brought on the 16th of Nisan; hence Shavuot always fell on or about the 6th of Sivan. Jewish sectarians, such as the Sadducees, rejected the rabbinic tradition concerning the date of the sheaf ceremony, preferring a later date, and celebrated Shavuot accordingly.

In Temple times, aside from the daily offerings, festival offerings, and first-fruit gifts, a special cereal offering consisting of two breads prepared from the new wheat crop was offered at the Temple. Since the destruction of the Second Temple, Shavuot observances have been dominated by its commemorative aspect. Many Jews spend the entire Shavuot night studying Torah, a custom first mentioned in the *Zohar* ("Book of Splendour"), a Kabbalistic work edited and published in the 13th–14th centuries. Some prefer to recite the *tiqqun lel Shavu'ot* ("Shavuot night service"), an anthology of passages from Scripture

Popperfoto

Family from Yemen celebrating Passover in Israel.

and the Oral Law (Mishna) compiled in the late medieval period. An expanded liturgy includes Hallel, public readings from the Torah, *yizkor* (in many congregations), and *musaf.* The Book of Ruth is read at the synagogue service, possibly because of its harvest-season setting.

Sukkot (Tabernacles)

Sukkot ("booths"), an ancient harvest festival that commemorates the booths the Israelites resided in after the Exodus, was the most prominent of the three pilgrim festivals in ancient Israel. Also called Ḥag ha-Asif (Festival of Ingathering), it has retained its joyous, festive character through the ages. It begins on Tishri 15 and is celebrated for seven days. The concluding eighth day (plus a ninth day in the Diaspora), Shemini Atzeret, is a separate holiday. In Temple times, each day of Sukkot had its own prescribed number of sacrificial offerings. Other observances, recorded in the Mishna tractate *Sukka,* include the daily recitation of Hallel, daily circumambulation of the Temple altar, a daily water libation ceremony, and the nightly *bet ha-sho'eva* or *bet ha-she'uvah* ("place of water drawing") festivities starting on the evening preceding the second day. The last mentioned featured torch dancing, flute playing, and other forms of musical and choral entertainment.

Ideally, Jews are to reside in booths—walled structures covered with thatched roofs—for the duration of the festival; in practice, most observant Jews take their meals in the *sukka* ("booth") but reside at home. A palm-tree branch (*lulav*), bound up together with myrtle (*hadas*) and willow (*'arava*) branches, is held together with a citron (*etrog*) and waved. Medieval exegetes provided ample (if not always persuasive) justification for the Bible's choice of these particular branches and fruit as symbols of rejoicing. The numerous regulations governing the *sukka, lulav,* and *etrog* comprise the major portion of the treatment of Sukkot in the codes of Jewish law. The daily Sukkot liturgy includes the recitation of Hallel, public readings from the Torah, the *musaf* service, and the circumambulation of the synagogue dais. On the last day of Sukkot, called Hoshana Rabba (Great Hoshana) after the first words of a prayer (hoshana, "save us") recited then, seven such circumambulations take place. Kabbalistic (mystical) teaching has virtually transformed Hoshana Rabba into a solemn day of judgment.

Hoshana Rabba is followed by Shemini Atzeret (Eighth Day of Solemn Assembly), which is celebrated on Tishri 22 (in the Diaspora also Tishri 23). None of the more distinctive Sukkot observances apply to Shemini Atzeret; but Hallel, public reading from the Torah, *yizkor* (in many congregations), *musaf,* and a prayer for rain in the Holy Land are included in its liturgy. Simḥat Torah (Rejoicing of the Law) marks the annual completion of the cycle of public readings from the Torah. The festival originated shortly before the gaonic period (*c.* 600–1050 CE) in Babylon, where it was customary to conclude the public readings annually. In Palestine, where the public readings were concluded approximately every three years, Simḥat Torah was not celebrated annually until after the gaonic period. Israeli Jews celebrate Simḥat Torah and Shemini Atzeret on the same day; in the Diaspora, Simḥat Torah is celebrated on the second day of Shemini Atzeret. Its joyous celebrations bring the Sukkot season to an appropriate close.

Ten Days of Penitence. The Ten Days of Penitence begin on Rosh Hashana and close with Yom Kippur. Already in Talmudic times they were viewed as forming an especially appropriate period of introspection and repentance. Penitential prayers (*seliḥot*) are recited prior to the daily morning service and, in general, during the period scrupulous observance of the Law is expected.

Rosh Hashana (New Year)

According to Mishnaic teaching, the New Year festival ushers in the Days of Judgment for all of mankind. Despite its solemnity, the festive character of Rosh Hashana is in no way diminished. In Scripture it is called "a day when the horn is sounded"; in the liturgy "a day of remembrance." In the land of Israel and in the Diaspora, Rosh Hashana is celebrated on the first two days of Tishri. Originally celebrated by all Jews on Tishri 1, calendrical uncertainty led to its being celebrated an additional day in the Diaspora and, depending upon the circumstances, one or two days in Palestine. After the calendar was fixed in 359, it was regularly celebrated in Palestine on Tishri 1 until the 12th century, when Provençal scholars introduced the two-day observance. Considerable speculation in recent literature concerning the origin of the Jewish New Year festival proves mostly that its early history can only be conjectured, not reconstructed.

Scroll of the Law being shown to the congregation on Simḥat Torah in a synagogue on the Tunisian island of Jazīrat Jarbah.

BBC Hulton Picture Library

The most distinctive Rosh Hashana observance is the sounding of the ram's horn (*shofar*) at the synagogue service. Medieval commentators suggest that the blasts acclaim God as Ruler of the universe, recall the divine revelation at Sinai, and are a call for spiritual reawakening and repentance. An expanded New Year liturgy stresses God's sovereignty, his concern for man, and his readiness to forgive those who repent. On the first day of Rosh Hashana (except when it falls on the Sabbath) it is customary for many to recite penitential prayers at a river, symbolically casting their sins into the river; this ceremony is called *tashlikh* ("thou wilt cast"). Other symbolic ceremonies, such as eating bread and apples dipped in honey, accompanied with prayers for a "sweet" and propitious year, are performed at the festive meals.

Yom Kippur (Day of Atonement)

The most solemn of the Jewish festivals, Yom Kippur is a day when sins are confessed and expiated and man and God are reconciled. It is also the last of the Days of Judgment and the holiest day of the Jewish year. Celebrated on Tishri 10, it is marked by fasting, penitence, and prayer. Work, eating, drinking, washing, anointing one's body, sexual intercourse, and donning leather shoes are all forbidden.

In Temple times, Yom Kippur provided the only occasion for the entry of the high priest into the Holy of Holies; details of the expiatory rites performed by the high priest and others are recorded in the Mishna and recounted in the liturgy. Present-day observances begin with a festive meal shortly before Yom Kippur eve. The Kol Nidre prayer (recited before the evening service) is a legal formula which absolves Jews from fulfilling solemn vows, thus safeguarding them from accidentally violating a vow's stipulations. The formula first appears in gaonic sources (derived from the Babylonian Talmudic academies, 6th–11th centuries) but may be older; the haunting melody that accompanies it is of medieval origin. Virtually the entire day is spent in prayer at the synagogue, the closing service (*ne'ila*) concluding with the sounding of the ram's horn.

Minor festivals: Ḥanukka and Purim. Ḥanukka and Purim are joyous festivals lacking the work restrictions characteristic of the major festivals.

Ḥanukka (Feast of Dedication)

Ḥanukka commemorates the Maccabean (or Hasmonean) victories over the forces of the Seleucid king Antiochus IV Epiphanes (reigned 175–164 BCE), and the rededication of the Temple Kislev 25, 164 BCE. Led by Mattathias and his son Judah Maccabee, the Maccabees were the first Jews who fought to defend their religious beliefs rather than their lives. Ḥanukka is celebrated for eight days beginning on Kislev 25. The Ḥanukka lamp or candelabrum (*menora*), which recalls the Temple lampstand, is kindled each evening. One candle is lit the first evening; an additional candle is lit each subsequent evening until eight candles are lit on the last evening. According to the Talmud (Shabbat 21b), the ritually pure oil available at the rededication of the Temple was sufficient for only one day's light but miraculously lasted for eight days, hence the eight-day celebration of Ḥanukka. Evidence from the Apocrypha (writings excluded from the Jewish canon but included in the Roman Catholic and Eastern Orthodox canons) and rabbinic literature shows an association between Sukkot and Ḥanukka, possibly accounting for the latter's eight-day duration. Ḥanukka joy is expressed in festive meals, song, games, and gifts to children. The liturgy includes Hallel, public readings from the Torah, and the *'al ha-nissim* ("for the miracles") prayer. The Scroll of Antiochus, an early medieval account of Ḥanukka, is read in some synagogues and homes.

Purim (Feast of Lots)

As recorded in the biblical Book of Esther, Purim commemorates the delivery of the Persian Jewish community from the plottings of Haman, Ahasuerus' (perhaps Xerxes, king of Persia, 486–465 BCE) prime minister. Mordecai and his cousin Esther, the King's Jewish wife, interceded on behalf of the Jewish community, rescinded the royal edict authorizing a pogrom against the Jews, and instituted the Purim festival. The historicity of the biblical account is questioned by many modern scholars. It is now generally conceded that the Book of Esther was written in the Persian period (it contains Persian but not Greek words) and reflects Persian custom. Except for the Book of Esther, the earliest mention of the Purim festival is from the 2nd–1st centuries BCE. The name of the festival was derived from the Akkadian *pûru*, meaning "lot."

In most Jewish communities, Purim is celebrated on Adar 14 (some also celebrate it on the 15th, others only on the 15th). On the evening preceding Purim, men, women, and children gather in the synagogue to hear the Book of Esther read from a scroll (*megilla*). The reading is repeated Purim morning. A festive meal during the day is accompanied by much song, wine, and merriment. Masquerades, Purim plays, and other forms of parody are common. Friends exchange gifts of foodstuffs and also present gifts to the poor. Aside from the Esther readings, the liturgy includes public reading from the Torah and recital of the Purim version of the *'al hanissim* prayer.

The five fasts. The commemorative aspects of the fasts are bound up with their penitential aspects, all of which find expression in the liturgy. Thus the Jew not only relives the tragic history of his people with each fast, but is also afforded an opportunity to search within himself and focus on his own (and his people's) present and future. Penitential prayers (*seliḥot*) are recited on all fasts, and the Torah is read at the morning and afternoon services.

'Asara be-Ṭevet (Fast of Ṭevet 10) commemorates the beginning of the siege of Jerusalem by Nebuchadnezzar, king of Babylonia, in 588 BCE.

Shiva' 'Asar be-Tammuz (Fast of Tammuz 17) commemorates the first breach in the wall of Jerusalem by the Romans in 70 CE. It initiates three weeks of semi-mourning that culminate with Tisha be-Av.

Tisha be-Av (Fast of Av 9) commemorates the destruction of the First and Second Temples in 586 BCE and 70 CE. The most solemn of the five fasts, its self-denials are more rigorous than those prescribed for the others, and, like Yom Kippur, the fast begins at sunset. The book of Lamentations is read at the evening service, followed by poetic laments that are also recited Tisha be-Av morning.

Tzom Gedaliahu (Fast of Gedaliah) commemorates the slaying of Gedaliah, governor of Judah after the destruction of the First Temple.

Ta'anit Esther (Fast of Esther), which commemorates Esther's fast (*cf.* Esther 4:16), is first mentioned in gaonic literature.

The lesser holidays. A major festival in the biblical period, Rosh Ḥodesh (First Day of the Month) gradually lost most of its festive character. Since Talmudic times, it has been customary to recite Hallel on Rosh Ḥodesh. In the medieval period, aside from the liturgical practices carried over from the Talmudic period, it was celebrated with a festive meal. Always more diligently observed in Palestine than in the Diaspora, attempts to revive its full festive character are being made in modern Israel.

First mentioned in the Mishna, where it marks the New Year for tithing purposes, Tu bi-Shevat (New Year for Trees) assumed a festive character in the gaonic period, and later in the medieval period it became customary to eat assorted fruits on the holiday. In modern times it is associated with the planting of trees in Israel.

Lag ba-'Omer (33rd Day of the 'Omer Counting) is a joyous interlude in the otherwise somber period of 'Omer counting (*i.e.*, of the 49 days to Shavuot), which is traditionally observed as a time of semi-mourning. Usually celebrated as a school holiday with outings, it is first mentioned in medieval sources, which attribute its origin to the cessation of a plague that was decimating the students of Akiba, an influential rabbinic sage in the 2nd century, and to the anniversary of the death of another great rabbi, Simeon ben Yoḥai (died *c.* 170 CE).

The situation today. Modern attitudes toward the Sabbath and festivals vary considerably. Acculturated Jews under the sway of Western secularism often are ignorant of, or choose to neglect, traditional observances. Attitudes of committed Jews in the Western world are patterned mostly along the lines of accepted Orthodox, Conservative, and Reform practice. Thus for example, driving to synagogue services on the Sabbath is unthinkable in Orthodox circles, a matter of dispute among Conservative rabbis, and normative practice for Reform Jews. Among Orthodox Jews, who best preserve the traditional observances, contemporary discussion centres mostly on technological advances and their effect on Halakhic practice (the behaviour laid down in the written and oral Torah). Whether or not hearing aids may be worn on the Sabbath, and how crossing the international dateline affects observance of Sabbaths and festivals typify the sort of problem raised in Orthodox *responsa* ("replies" to questions on law and observance). Recent (and often heated) discussion in Conservative literature raises the possibility of abolishing the obligatory character of the additional festival days in the Diaspora (except for the second day of Rosh Hashana), thus unifying Jewish practice throughout the world. Reform Jews, the most innovative of the three groups, observe neither the additional festival days (including the second day of Rosh Hashana) nor the fasts and have introduced numerous modifications in the liturgy as well as in the observances. In recent years more radical Reform congregations have experimented freely with "psychedelic" sound and light effects and other novel forms of synagogue service.

In Israel Sabbath is the national day of rest, and Jewish holidays are vacation periods. Municipal ordinances govern public observance of the Sabbath and festivals; their enactment and enforcement vary with the political influence of the local Orthodox Jewish community. Attempts to interpret festivals along nationalistic lines are common; some kibbutzim (communal farms) stress the agricultural significance of the festivals. Independence Day is a national holiday; the preceding day, Remembrance Day, commemorates Israel's war dead. Yom Hashoa (Holocaust Day)—marking the systematic destruction of European Jewry between 1933 and 1945 and recalling the short-lived Ghetto uprisings—is commemorated officially on Nisan

27; many religious Israelis prefer to commemorate it on Ṭebet 10 (a fast day) now called *yom ha-qaddish* (day upon which the mourner's prayer is recited). Since the June 1967 war, Iyyar 28—Liberation of Jerusalem Day—is celebrated unofficially by many Israelis. Appropriate services are conducted on all the aforementioned holidays by most segments of Israel's religious community.

In Israel and the Diaspora, Jewish theologians often stress the timelessness and contemporaneity of holiday observances. Nevertheless, "revised" Passover Haggadot (plural of Haggada) in which contemporary issues are accorded a central position, appear regularly.

Scholarly research into the origin of the festivals, if unabated, has not advanced significantly in recent years, nor is it likely to unless new evidence is forthcoming. Attempts to trace the development and spread of festival observances have fared better, and studies such as A. Yaari's *History of the Simḥat Torah Festival* (in Hebrew) bode well for the future. (S.Z.L.)

HOLY PLACES: THE LAND OF ISRAEL AND JERUSALEM

The land of Israel, as is evident from the biblical narratives, played a significant role in the life and thought of the Israelites. It was the promised home, for the sake of which Abraham left his birthplace; the haven toward which moved the tribes who escaped from Egyptian servitude; the hope of the exiles in Babylon. In the long centuries following the destruction of the Judean state by the Romans, it remained inextricably bound up with messianic and eschatological expectations. During the early period of settlement, there seem to have been many sacred localities, with one or another functioning for a time as a central shrine for all of the tribes, without displacing the others. Even the establishment of Jerusalem as the political capital by David and the building of a royal chapel there by Solomon did not bring to an end local cult centres. It was not until the reign of Josiah of Judah (640–609) that a reform centralized the cult in Jerusalem and attempted—although not entirely successfully—to end worship at local shrines. However irregular was the effectiveness of this reform, the Babylonian Exile and the subsequent return saw Jerusalem and its Temple win out over its rivals and become—in law, in fact, and in sentiment—the centre of Jewish cultic life. As noted above, this did not inhibit the rise and development of other forms of worship and even—on a few occasions—other cult centres. Nonetheless, no matter how unpopular the priesthood of the Jerusalem Temple became with some segments of the population—the Qumrān community seems to have denied its legality, and the Pharisees complained bitterly about its arrogance and exactions, attempting when politically feasible to impose and enforce Pharisaic regulations upon it—reverence for the Temple itself seems to have remained a widespread sentiment. With the destruction of the Temple by the Romans in 70 CE, such reverence was transformed both by messianic expectations and eschatological hopes into fervent devotion, which, over the following centuries, became idealized and even supernaturalized. The most ardently articulated statement of the crucial role of the land of Israel and the Jerusalem Temple is found in the *Sefer ha-Kuzari* of Judah ha-Levi in which the two are seen as absolutely indispensable for the proper relation between the people Israel and God. Symbolizing the significance of the land and of the city is the practice of facing in their direction during worship. The earliest architectural evidence derived from synagogue remains in Galilee indicates that the attempt was made to arrange the building in such a way that the worshippers faced directly toward Jerusalem. This practice may have continued even in the Diaspora, but at a later date the present practice of setting the holy ark in or before the east wall was established, so that "facing Jerusalem" is now more symbolic than actual.

The Jerusalem Temple

THE SACRED LANGUAGE: HEBREW AND THE VERNACULAR TONGUES

The transformation of Hebrew into a sacred language is, of course, bound up with the political fate of the people. In the period following the return from the Babylonian Exile, Aramaic, a cognate of Hebrew, functioned as the international or imperial language in official life and certainly gained a foothold as a vernacular. It did not, despite claims made by some scholars, displace the everyday Hebrew of the people. The language of the Mishna, far from being a scholar's dialect, seems to reflect—in the same way as the Koine (common) Greek of the New Testament—popular speech. Displacement of Hebrew—both in its literary form in Scriptures and in its popular usage—did take place in the Diaspora, however, as evidenced by the need to translate Scriptures into Greek in some communities and into Aramaic in others. As far as the emerging order of worship is concerned, there seems also to have been an inclination on the part of some authorities to permit even the recitation of the Shema complex in the vernacular. Struggles over these issues within the communities continued for a number of centuries in various places, but the development of formal literary Hebrew—a sacred tongue, to be used side by side with the Hebrew Scriptures in worship—brought them to an end. Although the communities of the Diaspora used the vernaculars of their environment in day-to-day living and even—as in the case of the communities of the Islāmic world—for philosophical, theological, and other scholarly writings, in worship, Hebrew remained the standard until modern times when some of the reform movements in western Europe sought partially—and a very small fraction even totally—to displace it.

Use of the vernacular in worship

THE RABBINATE

Legal, judicial, and congregational roles. The rabbinate, with its peculiar nature and functions, is the result of a series of developments going back to the period that followed the disastrous second revolt against Rome (132–135 CE). The term rabbi ("my teacher") was originally an honorific title for the graduates of the academy directed by the *nasi* or patriarch, the head of the Jewish community in Palestine in that era, who was also a Roman imperial official. The curriculum of the school was Torah, written and oral, according to the Pharisaic tradition and formulation. The *nasi* appointed rabbis to the law court (the Bet Din) and as legal officers of local communities: acting with the local elders, they supervised and controlled the life of the community and its members in all of its aspects. A similar situation obtained in Babylon under the Parthian and Sāsānian empires, where the *resh galuta* or exilarch ("head of the exile") appointed rabbinical officials to legal and administrative posts. In time the patriarchate and exilarchate disappeared, but the rabbinate, nourished by independent rabbinical academies, survived. An authorized scholar, when called to become the judicial officer of a community, would at the same time become the head of the local academy and would, after adequate preparation and examination, grant authorization to his pupils, who were then eligible to be called to rabbinical posts. There was, thus, a diffusion of authority, the communities calling, rather than a superior official appointing, their rabbis. What must be kept in mind is that these rabbis were not ecclesiastical personages but communal officials, responsible for the governance of the entire range of life of what was understood to be the *qehilla qedosha*, the "holy community."

In modern times and particularly in the Western world, the total change in Jewish communal existence required a transformation of this ancient structure. The rabbinate became, for the most part, an ecclesiastical rather than a communal agency, reflecting the requirements of civic life in modern national states. The education of rabbis who now function within this new situation is carried on in seminaries whose structure and curriculum have been influenced by European and American academic institutions. The majority of their graduates serve as congregational rabbis, in roles similar to those of ministers and priests in the Christian denominations, but with some other functions deriving from the particular situation and nature of the Jewish community.

The modern rabbi

Even in the State of Israel, where certain larger areas, such as that of family law, are still reserved to the rabbinate, it nonetheless functions more as a counterpart to other ecclesiastical organizations, Christian and Muslim, than as an overarching and all-inclusive communal agency that embodies, as in the past, involvement in every aspect of community and personal life.

Chief rabbinates. The existence of the offices of chief rabbi in the State of Israel derives from the situation in the Turkish Empire when the various religious communities functioned as quasi-political entities in that multiethnic conglomerate. Israel has two chief rabbis, one for the Ashkenazic (European) and one for the Sefardic (Oriental) communities—they no longer function, however, as the heads of whole communities but only of ecclesiastical organizations. The same is true in those countries outside Israel that have the office of chief rabbi; *e.g.*, Great Britain and France. Here they function vis à vis the governments like their ecclesiastic counterparts in the Christian churches. While they have certain kinds of limited authority because of their official position, their jurisdiction extends only over those members of the total Jewish community who are ready to accept it; others form their own ecclesiastical units and act without reference to the chief rabbinate. In some situations, particularly in the United States where there is no similar structure, the title chief or grand rabbi has been assumed occasionally by individuals as the means of asserting superior dignity or even (fruitlessly) authority.

GENERAL COUNCILS OR CONFERENCES

The precise nature of the Sanhedrin (Council Court) in the last years of the Jewish commonwealth is a much disputed matter. The several councils mentioned in Talmudic literature are equally difficult to define with exactitude. There are references scattered throughout medieval literature that suggest the existence of councils and synods but their composition and authority are also uncertain. Around 1000 a synod was held in the Rhineland in which French and German communities participated under the guidance of Rabbenu Gershom, the leading rabbinic authority of the region. The late Middle Ages saw the rise in eastern Europe of the Waʿad Arbaʿ Aratzot (Council of the Four Lands) composed of communal representatives from Great Poland, Little Poland, Russian Poland (Volhynia), and Lithuania. At the beginning of the modern era Napoleon (1806) summoned an Assembly of Notables—representatives of communities under French dominion—to deal with questions arising from the dissolution of the older status of the Jews and their naturalization as individuals into the new national states. Those decisions of the Assembly that involved questions of Jewish law were subsequently submitted to a Grand Sanhedrin called into being by Napoleon to provide some sort of Halakhic justification for the acts the French imperial government had required of the Jewish communities. During the 19th century the demand for the reform of Jewish life—principally the liturgy of the synagogue, but many other aspects as well—evoked a series of rabbinical conferences and synods that debated the questions and sought to guide the changes thought to be necessary. A similar procedure was followed on the American scene. In both instances, after an initial period in which radicals, moderates, and conservatives argued their respective cases in the same forum, polarization set in and intellectual differences were transformed into competing organizations. In the 1970s the several tendencies within the Jewish communities in North America were institutionalized in rabbinical conferences and congregational unions—Orthodox, Conservative, and Reform—whose influence was in large measure limited to their adherents. In the United States the Synagogue Council of America claims to be the "united voice" of American Jewry in common concerns. There is also a worldwide body in Reform or Liberal Judaism—the World Union for Progressive Judaism.

MODERN VARIATIONS

The above sketch of basic practices and institutions has attempted to describe the so-called traditional situation, although it has been indicated that even here there are variations—actually more than have been noted. In addition, reference has been made to some shifts and changes that represent a giving up of traditional practices on the basis of intellectual decisions about the nature of Judaism, its beliefs, practices, and institutions. Such changes are far too numerous to describe in detail. What is more important is to indicate their motivation. Basically, it is the view that the Halakhic system is not, as a whole and in all of its parts, divinely revealed but is rather a human process that seeks to expose in mutable forms the meaning of the divine-human encounter. Thus viewed, the practices and institutions are understood to be historically determined, reflecting the multifaceted experience of the people Israel as it has sought to live in the presence of God. Historical scholarship has, from this point of view, disclosed the origins, rise, development, and decline of these structures in the past and thus authorizes such changes in the present and future as appear to fulfill the needs of the community and its members. An examination of the specific deviations from the traditional forms makes clear that the application of this position, or attitude, has been subject to wide variation during the 19th and 20th centuries, in which it has operated. Some have seen it as a call for the disengagement from much if not all of the traditional pattern, and a recognition that only the spiritual essence is of importance or consequence for Judaism. Others have argued that an indiscriminate use of historicism (the explanation of values and forms in terms of their historical conditions) is unjustified and that the burden of proof is always upon those who would introduce changes. In the post-World War II period, the question has been whether a reconstituted Halakhic system might not be a requirement of the day.

Reinterpretations of the Halakhic system

Art and iconography

The anti-iconic principle and its modifications. Although the Second Commandment (Ex. 20:4; Deut. 5:8), "You shall not make yourself a graven image, or any likeness of anything that is in heaven above, or that is in the earth beneath, or that is in the water under the earth," has indeed been understood as absolutely prohibiting any and all artistic representation, this is not the only way in which these words may be interpreted. What is intended is a prohibition against the construction of such likenesses as were the object of worship in the cultural area in which the Israelites dwelt. Even in the Bible there are reports of artistic productivity in the construction of the tent sanctuary and its ritual vessels (Ex. 25–31) and of the Temple in Jerusalem (I Kings 6–7). The literalness and rigour with which the commandment was interpreted depended upon the larger situation of the community, so that during periods of external pressures toward religious conformity, such as the reign of Antiochus IV Epiphanes in Antioch (175–164 BCE), the anti-iconic attitude sharpened. Similarly, during the Roman occupation, the presence of the battle standards of the legions with their animal representations was looked upon as an affront, while extreme Pietists would not even handle Roman coinage because of the images stamped on it. On the other hand, the walls of a 3rd-century-CE synagogue in Doura-Europus in Syria are covered from floor to ceiling with biblical scenes with human representations, and a number of synagogues in Palestine had elaborate mosaic floors with the signs of the zodiac, representations of the seasons, and the like. Further, illuminated manuscripts from the medieval period in Europe were frequently decorated with biblical figures, some quite clearly copied from Christian prototypes. A fascinating mediating position is to be seen in a Haggada, in which the human figures have bird heads. Synagogues from a later, although preemancipation, period (before the 18th century) were often decorated with animal figures. In the modern period the avoidance of human figures has not been entirely accomplished, although nothing like the decorations of Doura-Europus has appeared.

The synagogue at Doura-Europus

Ceremonial objects and symbols. Nonetheless, given this general anti-iconic attitude, much of Jewish artistic endeavour has been directed toward the creation of ceremonial objects: Qiddush goblets, candlesticks and candelabra, spice boxes for the Havdala ceremony at the end of the Sabbath, ornamented containers for the *mezuza* (a

Sacrifice of Isaac, detail of a mosaic from the synagogue of Bet Alfa, Jezreel, Israel, 6th century AD.

Picture from the photographic archive of the Jewish Theological Seminary of America, New York. Frank J. Darmstaedter.

parchment on which is written Deut 6:4–9 and 11:13–21, fastened to the doorpost on the right side as one enters), the silver crowns placed on the Torah scrolls, together with the mantles and breastplates for the same, and many other objects designed to embellish the performance of the large number of ritual acts of the individual and the community. All of these vary in artistic quality, from the work of simple artisans to exquisitely produced works of master craftsmen.

Architecture. The building of synagogues, too, is an expression of artistic interest and concern, as well as of religious and social function. Nothing is known of these edifices, if indeed there were any, until the Greco-Roman period. Then the Roman basilica often provided the appropriate model. What was required was a spacious hall for assembly, and galleries for the women, and this form served that purpose very well. However synagogues were furnished before the destruction of the Second Temple, after that event some attempt seems to have been made to transfer some of the latter's appurtenances to the former, a move that was successfully resisted. When possible, the synagogue stood on a hill. Before it stood a walled entrance court with a fountain for ablutions. Before the Temple destruction, the building may have been oriented with its doors facing eastward, but afterward they faced Jerusalem; still later, when the holy ark containing the Torah scrolls was placed in a fixed position, the orientation was reversed so that the central gate would not be blocked; ultimately, the ark was placed in or against the east wall, without reference to the actual direction of Jerusalem. As the Diaspora grew larger, the new communities adapted the architectural forms of the enveloping culture. The surviving buildings of the Muslim period in Spain are often built with the horseshoe arches and decorated with the exquisite stucco arabesques that mark the era. The medieval period in Christian Europe saw a revival of a very strict anti-iconic attitude and a gradual rejection of the church edifice in favour of secular buildings as a model for the synagogue. The increasingly limited role of the Jew in that society and the enlargement of restrictions by church and state made it necessary to modify the synagogal structure. The doors no longer were in the wall facing the ark; the courtyard grew smaller; galleries were discontinued (side rooms now serving as the women's section); and a double- rather than a triple-aisled construction was largely favoured. Similar developments took place in eastern Europe with the building of fortress-synagogues and the remarkable wooden synagogues of Poland. In the early postemancipation period, Baroque style had its day, followed by Greek temples, Romanesque, Gothic, and pseudo-Byzantine churches, and pseudo-Moorish mosques. In the most recent period, the various schools of functionalism and their commercial descendants have come to the fore. The best of these have brought together fine architectural design and beautifully conceived and executed decoration. The interior arrangement, even in some traditional synagogues, has been influenced by the Protestant sermon-centred form of worship, so that some of the unique forms that marked older structures are absent. The holy ark is, however, still a centre of attention and has often been treated in interesting and striking ways.

Synagogal sites and structures

Paintings and illustrations. As noted above, the use of paintings in the decoration of synagogues goes back to at least the 3rd century CE and is found in the late pre-emancipation and modern synagogues as well. Manuscripts, too, were illuminated with miniatures and the Renaissance period saw the appearance of beautifully decorated Scrolls of Esther and *ketubbot* (marriage contracts). Nonetheless, the appearance of Jewish artists in painting and sculpture is a modern phenomenon. Beginning in the 19th century, interest grew apace and more and more Jews are to be found, often in the avant-garde of these fields. Some, such as Marc Chagall and Jacques Lipchitz, have done specifically religious art.

Music. The description of the synagogue service above noted the role of the *ḥazzan,* or cantor. It is he who reads the service and declaims the scriptural lessons to certain set musical modes that vary with the season and occasion. Many of these call for melodic responses on the part of the congregation. The origins and varying developments

Liturgical cantillation

Picture from the photographic archive of the Jewish Theological Seminary of America, New York. Frank J. Darmstaedter

Pewter Passover plate, German, 17th century. Adam and Eve are represented in the centre, surrounded by the signs of the zodiac and a depiction of the sacrifice of Isaac. The Hebrew lettering on the rim indicates the order of the Passover eve service. In the Jewish Museum, New York.

of these chants are ancient, often obscure, and equally complicated. Whatever the basic materials, these were enlarged, varied, corrupted, and reworked over the centuries in the various environments in which the Jewish communities have lived. In modern times musicologists have begun to examine with great care the history of synagogal music, analyzing its basic structures and its relationship to the music of Christian liturgical traditions. In the 19th century in Western Europe much of the traditional music was either discarded or re-worked under the influence of western forms and styles. In addition the pipe-organ was introduced and was the centre of stormy controversy.

Literature. Literature has been throughout the ages the home of Jewish artistic activity. The Hebrew Bible is a work of monumental artistry, exhibiting grandeur of form and language in historical narrative, poetry, rhetoric, and aphorism. The extrascriptural writings of the period, although their originals have often vanished, still disclose literary genius of a high order in translation. The documents of the rabbinic tradition are not often looked at with an eye to their literary worth but much of the material, particularly the Haggadic portions of the Midrashim, reveals a noteworthy sensitivity to the uses of language. In the medieval period much attention was given to the production of *piyyuṭim*, liturgical poetry with which to embellish the *Siddur* (prayer book), itself a collection containing much imaginative, as well as pedestrian, writing. In the Islāmic world, under the influence of Arabic poetry, Hebrew poetry rose to a high peak in both liturgical and secular forms. The Middle Ages in the Rhineland also saw the beginnings of the Jewish form of Middle High German that was, over the centuries, to develop into an autonomous Jewish language, Yiddish, which, in the 19th century, became a literary vehicle of very high order. The same period saw the beginnings of the recreation of Hebrew into a literary language that has become the basis of the spoken vernacular of the State of Israel and of a flourishing literature encompassing every branch of the field. Since the emancipation at the end of the 18th century, Jews in western Europe and later in the United States have turned to literature in the vernaculars of their countries, and have produced writers of note dealing with both Jewish and general themes. (L.H.S.)

Jewish philosophy

The term Jewish philosophy refers to various kinds of reflective thought engaged in by persons identified as being Jews, in one sense or another. At times, as in the Middle Ages, this meant any methodical and disciplined thought, whether on general philosophical subjects or on specifically Judaic themes, pursued by Jews. In other eras, as in modern times, concentration on the latter has been considered a decisive criterion, so that philosophers who are Jewish but unconcerned with Judaism or the Jewish heritage and destiny in their thought are not ordinarily classified as Jewish philosophers.

PRE-HELLENISTIC AND HELLENISTIC THOUGHT

Bible and Apocrypha. Philosophy arose in Judaism under Greek influence; however, a kind of philosophical approach may be discerned in early Jewish religious works apparently subject to little or no Greek influence. The books of Job and Qohelet (Ecclesiastes) were favourite works of medieval philosophers, who took them as philosophical discussions untinged by theological preconceptions. The book of Proverbs introduces, in an apparently theological context, the concept of Wisdom (Ḥokhma), which was to have a primordial significance for Jewish philosophical and theological thought, and presents it as the first and favourite of God's creations. It is also praised in the book of the Wisdom of Jesus the Son of Sirach (Ecclesiasticus) as instilled by God into all his works and granted in abundance to those he loves. It is sometimes equated with fearing God and keeping the Law; however, in other passages piety seems to be regarded as superior to Wisdom. The Wisdom of Solomon, probably originally written in Greek, praises Wisdom, which is held to be an image of God's goodness and a reflection of the eternal light. God is said to have given the author knowledge of the composition of the world, of the powers, the elements, the nature of animals, the divisions of time, and the positions of the stars. In its vocabulary and perhaps in some of its doctrines, the work shows the influence of Greek philosophical conceptions. It has had considerable influence on Christian theology.

The concept of Wisdom

Philo Judaeus. The first systematic attempt to apply Greek philosophical concepts to Jewish doctrines was made by Philo Judaeus (Philo of Alexandria) in the 1st century CE. Philo, a scholar who combined Greek and Jewish learning, was influenced by Platonic and Stoic writings, and probably also by certain postbiblical Jewish beliefs and speculations. He apparently had some knowledge of the Oral Law, which was being evolved in his time, and he also knew of the Essenes, a contemporary rigorous sect, whom he praised highly.

Philo's main contribution was interpretative. He provided Jewish conceptions with the hallmark of intellectual and cultural respectability by stating them in Greek philosophical terms; he also showed that many Greek notions were consonant with Jewish doctrine, as he conceived it, and with the allegorical sense of biblical texts, as he read them. He had two schemes of reference—Jewish religious tradition and Greek philosophy—and the fact that he took care to stress the primacy of the former may have been more than mere lip service. It may be argued with some plausibility that in central points of his thought, such as his conception of the Logos (the divine Reason or Word), Philo used philosophical notions as trappings for an originally religious belief. A main function of the Logos in his thought is to serve as an intermediary between the transcendent, unknowable God and the world, a view that probably has a close connection with the view of his Jewish contemporaries concerning the world of God, by means of which he accomplishes his designs. On basic philosophical or theological problems, such as the creation of the world or the freedom of will, Philo's writings provide either vague or contradictory answers. He placed mystic ecstasy, of which he may have had personal experience, above philosophical and theological speculations.

The concept of Logos

Philo's approach, his method of interpretation, and his way of thinking, as well as some of his conceptions—primarily that of the Logos—exerted a considerable influence on early Christian thought but not, to any comparable extent, upon Jewish thought in that period. Later, in the Middle Ages, knowledge of Philo among Jews was either very slight or nonexistent. Not until modern times was his role in Jewish religious thought recognized.

Other ancient sources. Some traces of a knowledge of popular, mainly Stoic philosophy may be found in the Mishna, a codification of the Oral Law composed in Palestine in the 2nd century CE, and in the subsequent Talmudic literature set down in writing in Palestine and Babylonia. On the whole these traces are rather slight. Some scholars believe that the influence of Greek philosophy on Palestinian Jewry was far-reaching, but the case, to say the least, is not proven. Jewish theological and cosmological speculations occur in the Midrashim (plural of Midrash,), which, under the guise of interpreting biblical verses, propound allegorical interpretations, legends, and myths, and in the *Sefer Yetzira* ("Book of Creation"), a work that is a combination of a cosmogony and a grammar and that was fictitiously attributed to Abraham. There is no clear evidence of the period in which it was written; both the 3rd century and the 6th or 7th century have been suggested. The book became a key work in later Jewish mysticism.

MEDIEVAL PHILOSOPHY

In the 9th and 10th centuries, after a long hiatus, systematic philosophy and ideology reappeared among Jews, a phenomenon indicative of their accession to Islāmic civilization. The evolution of Islām in the 9th and 10th centuries showed that Greek scientific and philosophic lore could be separated, at least to some extent, from its pagan associations and could be transformed into another language and another culture; it also tended to show that a culture in which the sciences and philosophy or the

Islāmic background

sciences and theology or both of these combinations were an indispensable part could be based upon a monotheistic prophetic religion that in all relevant essentials, including adherence to a basic religious law, was closely akin to Judaism. The question of whether philosophy is compatible with religious law (the answer being sometimes negative) constituted the main theme of the foremost medieval Jewish thinkers. From approximately the 9th to the 13th centuries Jewish philosophical and theological thought participated in the evolution of Islāmic philosophy and theology and manifested only in a limited sense a specifically Jewish character. Jewish philosophers showed no particular preference for philosophic texts written by Jewish authors over those composed by Muslims, and in many cases the significant works of Jewish thinkers constitute a reply or a reaction to the ideas of Islāmic philosophic and scientific writings.

Jewish kalām. Although several Jewish intellectuals in 9th–10th-century Babylonia were steeped in Greek philosophy, the most productive and influential Jewish thinkers of this period represented a very different tendency, that of the Muʿtazilite *kalām*. *Kalām* (literally, "speech") is an Arabic term used both in Islāmic and in Jewish vocabulary to designate several theological schools that were ostensibly opposed to Greek, particularly Aristotelian, philosophy. The Aristotelians, both Islāmic and Jewish, regarded *kalām* theologians (called the *mutakallimūn*) with a certain contempt, holding them to be mere apologists, watchdogs of religion, and indifferent to truth. Herein they did not do justice to their adversaries, for many representatives of the Muʿtazilite school of *kalām*, formed in the 8th century, displayed a genuine speculative impulse. Its theology, forged in disputes with Zoroastrians, Manichaeans, and Christians, claimed to be based on reason.

Saʿadia ben Joseph. This belief in reason, as well as some of the tenets of Muʿtazilite theology, were taken over by Saʿadia ben Joseph, who was also influenced, either directly or through the intermediary of an Arabic philosopher, by the arguments of a Christian 6th-century philosopher, John Philoponus, against certain Aristotelian and Neoplatonic positions. Saʿadia's main theological work, *Kitāb al-amānāt wa al-iʿtiqādāt (Beliefs and Opinions)*, is modelled on similar Muʿtazilite treatises and on a Muʿtazilite classification of theological subject matter known as the Five Principles. Like many Muʿtazilite authors, Saʿadia starts out by setting forth in his introduction a list and theory of the various sources of knowledge.

Saʿadia's four sources of knowledge

Saʿadia distinguished four sources: (1) the five senses, (2) the intellect, or reason, (3) necessary inferences, and (4) reliable information given by trustworthy persons. In Saʿadia's sense of the word, intellect, or reason (*al-ʿaql*), means first and foremost an immediate, a priori cognition, independent of sense experience. In *Beliefs and Opinions* the intellect is characterized as having immediate ethical cognitions—that is, as discerning what is good and what is evil—in opposition to the medieval Aristotelians, who did not regard even the most general ethical rules as a priori cognitions. The third source of knowledge comprises inferences of the type "if there is smoke, there is fire," which are based on data furnished by the first two sources of knowledge. The fourth source of knowledge is meant to validate the teachings of Scripture and of the religious tradition. Teachings of Scripture must be held true because of the trustworthiness of the men who propounded them. One of the main purposes of the work is to show that the knowledge deriving from the fourth source concords with that discovered by means of the other three, or, in other words, that religion and human reason agree.

Saʿadia opposed Aristotle's view that the natural order was eternal. He held, with other partisans of the Muʿtazilite *kalām*, that the demonstration of the temporal creation of the world must precede and pave the way for the proof of the existence of God the Creator. Given the demonstrated truth that the world has a beginning in time, it can be proved that it could have been produced only through the action of a creator. It can further be proved that there can have been only one Creator.

The theology of Saʿadia, like that of the Muʿtazilites, hinges on two principles: the unity of God and the principle of justice. The latter principle takes issue with the view (widespread in Islām and present also in Judaism) that the definition of what is just and what is good depends solely on God's will, to which none of the moral criteria found among men is applicable; according to this view, a revelation from God can convert an action now generally recognized as evil into a good action. Against this way of thinking, Saʿadia and the Muʿtazilites believed that being good and just or evil and unjust are intrinsic characteristics of human actions and cannot be changed by divine decree. The notions of justice and of good, as conceived by man, are binding on God himself. Since, according to Saʿadia, man has a priori knowledge of good and evil, just and unjust, the fact that human ethical judgments are valid for God means that man's ethical cognitions are also those of the Deity.

Divine justice and human ethical judgment

The function of religious law—of central importance in traditional Judaism and Islām—is to impose on man the accomplishment of good actions and to prohibit bad ones. Because Saʿadia believed that man has a priori knowledge of good and evil and that this knowledge coincides with the principles underlying the most important portions of the revealed law, he was forced to ask the question whether this law is not superfluous. He could, however, point out that, whereas the human intellect recognizes that certain actions—for instance, murder or theft—are evil, it cannot by itself discover the best possible definition of what constitutes a particular transgression, nor can it, on its own, determine the punishment appropriate for a transgression. On both points, Saʿadia asserted, the commandments of religious law give the best possible answers.

The commandments that accord with the behests of the human intellect were designated by Saʿadia as the intellectual, or rational, commandments. According to him, they include the duty of manifesting gratitude to the Creator for the benefits he has bestowed upon man. Saʿadia recognized that a considerable number of commandments—for instance, those dealing with the prohibition of work on the Sabbath—do not belong to this category. He held, however, that the obligation to obey them may be derived from the rational commandment that makes it incumbent upon man to be grateful to God, for such gratitude entails obedience to his orders.

The Karaites. Saʿadia's adoption of the rational Muʿtazilite theology was a part of his overall activity, directed toward the consolidation of rabbinical Judaism (based on the Mishna and Talmud), which was being attacked by the Karaites. This Jewish sect, founded by Anan ben David in the 8th century, rejected the authority of the Oral Law—that is, of the Mishna and the Talmud. In the 10th century and afterward, the Karaites accepted as their guides the Bible (Old Testament) and human reason, in the Muʿtazilite sense of the word. Their professed freedom from any involvement with postbiblical Jewish religious tradition facilitated a rational approach to theological doctrine. This approach led the Karaite authors to criticize their opponents, the adherents of rabbinical Judaism, for holding anthropomorphic beliefs based, in part, on texts of the Talmudic period. Karaite authors propounded, in conceptual terms, a theology of Jewish history in exile (*galut*). Life in exile is a diminished existence; nevertheless, the good or bad actions of the Jewish people (rather than their material strength or weakness) affect the course of history. Redemption may come when all Jews are converted to Karaism.

Karaite rationalism and atomism

The Karaites adopted, wholesale, Muʿtazilite *kalām*, including its atomism. The Muʿtazilite atomists held that everything that exists consists of minute, discrete parts. This applies not only to bodies but also to space, to time, to motion, and to the "accidents"—that is, qualities, such as colour—which the Islāmic and Jewish atomists regarded as being joined to the corporeal atoms (but not determined by them, as had been believed by the Greek Atomists). An instant of time or a unit of motion does not continue the preceding instant or unit. All apparent processes are discontinuous, and there is causal connection between their successive units of change. The fact that cotton put into fire generally burns does not mean that fire is a cause of burning; rather, it may be explained as a "habit," signifying that this sequence of what is often wrongly held to be cause and effect has no character of necessity. God's free will is the only agent of everything that occurs, with the exception of one category—human actions. These are causes that produce effects; for instance, a man who throws a stone at another man, who is then killed, directly brings about the latter's death. This inconsistency on the part of the theologians was necessitated by the principle of justice, for it would be unjust to punish a man for a murder that was a result not of his action but of God's. This grudging admission that causality exists in certain strictly defined and circumscribed cases was occasioned by moral, not physical, considerations.

Jewish Neoplatonism. *Isaac Israeli.* Outside Babylonia, philosophical studies were pursued by Jews in the 9th and 10th centuries in Egypt and in the Maghrib (northwest Africa). The outstanding figure was Isaac ben Solomon Israeli, an Egyptian-born North African who has been called "the first Jewish Neoplationist." In his philosophical works, such as the "Book of Elements" and the "Book of Five Substances," he drew largely upon a 9th-century Muslim popularizer of Greek philosophy, Abū Yūsuf Yaʿqūb al-Kindī, and also, in all probability, upon a lost pseudo-Aristotelian text. The peculiar form of Neoplatonic doctrine that seems to have been set forth in this text had, directly and indirectly, a considerable influence on medieval Jewish philosophy.

According to Israeli, God creates through his will and power. The two things that were created first were form, identified with wisdom, and matter, which is designated as the genus of genera (the classes of things) and which is the substratum of everything, not only of bodies, but also of incorporeal substances. This conception of matter seems to derive from the Greek Neoplatonists Plotinus and Proclus, particularly from the latter. In Proclus' opinion generality was one of the main criteria for determining the ontological priority of an entity (relative place in the levels of reality). Matter, because of its indeterminacy, obviously has a high degree of generality; consequently, it figures among the entities having ontological priority. According to the Neoplatonic view, which Israeli seems to have adopted, the conjunction of matter and form gives rise to the intellect. A light sent forth from the intellect produces the rational soul, and in its turn it gives rise to the vegetative soul.

Israeli's doctrine of prophecy seems to be the earliest Jewish philosophical theory attributing prophecy to the influence of the intellect on the imaginative faculty. According to Israeli, this faculty receives from the intellect spiritual forms that are intermediate between corporeality and spirituality. This explanation implies that these forms, "with which the prophets armed themselves," are inferior to purely intellectual cognitions.

Solomon ibn Gabirol. In essentials the schema of creation and emanation propounded by Isaac Israeli and his Neoplatonic source or sources was taken over by Solomon ibn Gabirol, a celebrated 11th-century Hebrew liturgical poet, who seems to have been the earliest Jewish philosopher of Spain. His chief philosophical work, "Fountain of Life," written in Arabic, has been preserved in full only in a 12th-century Latin translation entitled *Fons vitae.* This work, which makes no reference to Judaism or to specifically Jewish doctrines, is a didactic dialogue between a disciple and a master who teaches him true philosophical knowledge. Despite its prolixity and many contradictions, it is an impressive work. Few medieval texts so effectively communicate the Neoplatonic conception of the existence of a number of planes of being that differ according to their ontological priority, the derivative and inferior ones constituting a reflection in a grosser mode of existence of those that are prior and superior.

Ibn Gabirol's doctrine of divine will

A central conception in Ibn Gabirol's philosophy is concerned with the divine will, which appears to be both part of and separate from the divine essence. Infinite according to its essence, the will is finite in its action. It is described as pervading everything that exists and as being the intermediary between the divine essence and matter and form. Will was one of a number of traditional appellations applied in various medieval theologies to the entity intermediate between the transcendent Deity and the world or to the aspect of the Deity involved in creation. According to a statement in *Fons vitae,* matter derives from the divine essence, whereas form derives from the divine will. This suggests that the difference between matter and form has some counterpart in the Godhead and also that universal matter is superior to universal form. Some of Ibn Gabirol's statements seem to bear out the impression of superiority of universal matter; other passages, however, appear to imply a superiority of universal form.

Form and matter, whether they be universal or particular, exist only in conjunction. All things, with the sole exception of God, are constituted through the union of the two, the intellect no less than the corporeal substance. In fact, the intellect is the first being in which universal matter and form are conjoined. The intellect contains and encompasses all things. It is through the grasp of the various planes of being, through ascending in knowledge to the world of the intellect and apprehending what is above it—the divine will and the world of the Deity—that man may "escape death" and reach "the source of life."

Judah ha-Levi. Judah ben Samuel ha-Levi (*c.* 1075–*c.* 1141), also from Spain and a celebrated Hebrew poet, was the first medieval Jewish thinker who consciously and consistently based his thought upon arguments drawn from Jewish history. His views are set forth in an Arabic dialogue, *al-Hazari* (Hebrew *Sefer ha-Kuzari*), the full title of which is translated as "The Book of Proof and Argument in Defense of the Despised Faith." This work is usually called *Kuzari*; *i.e.*, "the Khazar."

The *Kuzari*

Basing his narrative on the historical fact that the Khazars (a Turkic-speaking people in Central Eurasia) were converted to Judaism (*c.* 740), ha-Levi relates that their King, a pious man who did not belong to any of the great monotheistic religions, dreamed of an angel, who said to him, "Your intentions are pleasing to the Creator, but your works are not." To find the correct way of pleasing God, the King seeks the guidance of a philosopher, a Christian, a Muslim, then, finally, after hesitating to have recourse to a representative of a people degraded by its historical misfortune, of a Jewish scholar who converts him to Judaism. The words of the angel heard in a dream may, in accordance with both religious and philosophical doctrine, be regarded as a kind of revelation. The use of this element of the story enabled ha-Levi to suggest that it is not the spontaneous activity of human reason that impels man to undertake the quest for the true religion; for this, one needs the gift of prophecy, or, at least, a touch of the prophetic faculty (or a knowledge of the revelations of the past).

The argument of the philosopher whose advice is sought by the King brings this point home. This disquisition is a brilliant piece of writing, for it lays bare the essential differences between the Aristotelian God, who is totally ignorant of and consequently wholly indifferent to human individuals, and the God of religion. Within the framework of philosophical doctrine, the angel's words are quite meaningless. Not only is the God of the philosophers, who is a pure intellect, not concerned with man's works, but the (cultural) activities, involving both mind and body, to which the angel clearly referred, cannot, from the philosophical point of view, either help or hinder man in the pursuance of the philosophers' supreme goal—the attainment of union with the active intellect, a "light" of the divine nature. This union was supposed to confer knowledge of all intelligible things. Thus, man's supreme goal was here held to be of a purely intellectual nature.

In opposition to the philosopher's faith, the religion of ha-Levi's Jewish scholar is based upon the fact that God may have a close, direct relationship with man, who is not conceived primarily as a being endowed with intellect. The postulate that God can have intercourse with a creature made of the disgusting materials that go into the composition of the human body is scandalous to the King and prevents his acceptance of the doctrine concerning prophecy, expounded by the Muslim sage (just as the extraordinary nature of the Christological dogmas deters him from adopting Christianity).

The Jewish scholar's position is that it is contemplation

Knowledge of God through the contemplation of Jewish history

not of the cosmos but of Jewish history that procures knowledge of God. Ha-Levi was aware of the odium attaching to the doctrine of the superiority of one particular nation; he held, however, that only this doctrine explains God's dealing with mankind, which, like many other things, reason is unable to grasp. The controversies of the philosophers serve as proof of the failure of human intelligence to find valid solutions to the most important problems.

Ha-Levi's dialogue was also directed against the Karaites. He shows the necessity and celebrates the efficacy of a blind, unquestioning adhesion to tradition, which the Karaites rejected. Yet, he expounds a theology of Jewish exile that seems to have been influenced by Karaite doctrine. According to ha-Levi, even in exile the course of Jewish history is not determined like that of other nations by natural causes, such as material strength or weakness; the decisive factors are the religious observance or disobedience of the Jews. The advent of Christianity and of Islām prepares the other nations for conversion to Judaism, an event that will occur in the eschatological period (at the end of history).

Other Jewish thinkers c. 1050–c. 1150. During the period comprising the second half of the 11th century and the first half of the 12th century, many other Jewish thinkers appeared in Spain. Bahya ben Joseph ibn Pakuda wrote one of the most popular books of Jewish spiritual literature, *Kitāb al-hidāyah ilā farāʾiḍ alqulūb* ("Guidance to the Duties of the Heart"), which combines a theology influenced by Saʿadia with a moderate mysticism inspired by the teachings of the Muslim Ṣūfīs (mystics). The commandments of the heart—that is, those relating to men's thoughts and sentiments—are contrasted with the commandments of the limbs—that is, the Mosaic commandments enjoining or prohibiting certain actions. Bahya maintained that both sets of commandments should be observed (thus rejecting the antinomian position), but made clear that first and foremost he was interested in the commandments of the heart.

Abraham bar Hiyya Savasorda, a mathematician, astrologer, and philosopher, outlined in *Megillat ha-megalle* ("Scroll of the Revealer") a view of Jewish history that is reminiscent of ha-Levi but does not emphasize to the same degree the uniqueness of that history; it is also set forth in much less impressive fashion. Living in Barcelona under Christian rule, Bar Hiyya wrote his scientific and philosophical treatises not in Arabic but in Hebrew. Hebrew was also used by Abraham ibn Ezra (died *c.* 1167), a native of Spain who travelled extensively in Christian Europe. His commentaries on the Bible contributed to the diffusion among the Jews of Greek philosophical thought, to which Ibn Ezra made many, although as a rule disjointed, references. His astrological doctrine had a strong influence on some philosophers.

A borderline case: Abū al-Barakāt

The last outstanding Jewish philosopher of the Islāmic East, Abū al-Barakāt al-Baghdādī (who died as a very old man sometime after 1164), also belongs to this period. As a borderline case he illustrates a certain indeterminacy in the definition of a Jewish thinker. An inhabitant of Iraq, he was converted to Islām in his old age (for reasons of expediency, according to his biographers). His philosophy appears to have had a strong impact on Islāmic thought, whereas its influence upon Jewish philosophy and theology is very hard to pin down and may be practically nonexistent. His chief philosophical work, *Kitāb al-muʿtabar* ("The Book of That Which Has Been Established by Personal Reflection"), has very few references to Jewish texts or topics. Abū al-Barakāt rejects Aristotelian physics completely; according to him, time is the measure of being, and not, as Aristotle taught, the measure of motion, and he replaces Aristotle's bidimensional concept of place with the tridimensional notion of space, the existence of which is independent of the existence of bodies.

Jewish Aristotelianism. With regard to the adoption of Aristotelianism, including systems that in many essentials stem from but also profoundly modify the pure Aristotelian doctrine, there is a considerable time lag between the Islāmic East, on the one hand, and Muslim Spain and the Maghrib, on the other.

Abraham ibn Daud. Abraham ibn Daud (12th century), who is regarded as the first Jewish Aristotelian of Spain, was primarily a disciple of Avicenna, the great 11th-century Islāmic philosopher. According to a not unlikely hypothesis, he may have translated or helped to translate some of Avicenna's works into Latin, for Ibn Daud lived under Christian rule in Toledo, a town that in the 12th century was a centre for translators. His historical treatises, written in Hebrew, manifest his desire to familiarize his coreligionists with the historical tradition of the Latin world, which at that time was alien to most of them. But his philosophical work, *Sefer ha-emuna ha-rama* ("Book of Sublime Faith"), written in 1161 in Arabic, shows few, if any, signs of Christian influence.

The doctrine of emanation, set forth in the "Book of Sublime Religion," describes in the manner of Avicenna the procession of the ten incorporeal intellects, the first of which derives from God. This intellect produces the second intellect, and so on. Ibn Daud questioned in a fairly explicit manner Avicenna's views on the way the second intellect is produced; his discipleship did not by any means spell total adherence. Ibn Daud's psychology was also, and more distinctively, derived from Avicenna. The argumentation leading to a proof that the rational faculty is not corporeal attempts to derive the nature of the soul from the fact of immediate self-awareness. Like Avicenna, Ibn Daud tended to found psychology on a theory of consciousness.

Ibn Daud often referred to the accord that, in his view, existed between philosophy and religious tradition. As he remarked, the "Book of Sublime Faith" was not meant to be read either by readers who, in their simplicity, are satisfied with what they know of religious tradition or by those who have a thorough knowledge of philosophy. It was intended for readers of one type only, those who, being, on the one hand, acquainted with the religious tradition and having, on the other, some rudiments of philosophy, are "perplexed." It was for the same kind of people that Maimonides wrote his *Guide of the Perplexed.*

The greatest medieval Jewish philosopher

Maimonides. Maimonides (Moses ben Maimon, 1135–1204), a native of Spain, is incontestably the greatest name in Jewish medieval philosophy, but his reputation is not derived from any outstanding originality in philosophical thought. Rather, the distinction of Maimonides, who is also the most eminent codifier of Jewish religious law, is to be found in the vast scope of his attempt, in the *Dalālat al-hāʾirīn* (*Guide of the Perplexed*), to safeguard both religious law and philosophy (the public communication of which would be destructive of the law) without suppressing the issues between them and without trying to impose, on the theoretical plane, a final, universally binding solution of the conflict.

As Maimonides made clear in his introduction to the *Guide*, he regarded his self-imposed task as perilous, and he therefore had recourse to a whole system of precautions destined to conceal his true meaning from the people who, lacking the necessary qualifications, might misread the book and abandon observance of the law. According to Maimonides' explicit statement, these precautions include deliberately contradictory statements meant to mislead the undiscerning reader. The apparent or real contradictions that may be encountered in the *Guide* are perhaps most flagrant in Maimonides' doctrine concerning God. There seems to be no plausible hypothesis capable of explaining away the differences between the following three views:

1. God has an eternal will that is not bound by natural laws. Through an act of his will, he created the world in time and imposed on it the order of nature. This creation is the greatest of miracles; only if it is admitted can other miracles, which interfere with the causally determined concatenations of events, be regarded as possible. The philosophers' God, who is not free to cut the wings of a fly, is to be rejected. This conception is in keeping with the traditional religious view of God and is avowedly adopted by Maimonides because failure to do so would undermine religion.

2. Man is incapable of having any positive knowledge concerning God. No positive attributes—*e.g.*, wisdom or life—can be ascribed to God. Contrary to the attributes

predicated of created beings, the divine attributes are strictly negative; they state what God is not: for instance, he is not not-wise, and such a statement is not a positive assertion. Hence, only a negative theology is possible—saying what God is not. The way God acts can, however, be known. This knowledge is to be found in natural science.

3. God is an intellect. The formula current among medieval philosophers that maintains that in God the knowing subject, the object known, and the act of intellectual knowledge are identical derives from Aristotle's thesis that God knows only himself. Maimonides, however, in adopting the formula, interpreted it in the light of human psychology and epistemology (theory of knowledge), pointing out that, according to a theory of Aristotle, the act of human (not only of divine) cognition brings about an identity of the cognizing subject and cognized object. The parallel drawn by Maimonides between the human and the divine intellect quite evidently implies a certain similarity between the two; in other words, it is incompatible with the negative theology of other passages of the *Guide*. Nor can it be reconciled with his theological doctrine that the structure of the world—created in time—came into being through the action of God's will.

Maimonides' doctrine of prophecy

The enigma of the *Guide* would be nonexistent if Maimonides could be held to have believed that truth can be discovered in a suprarational way, through revelations vouchsafed to the prophets. This, however, is not the case. Maimonides held that the prophets (with the exception of Moses) combine great intellectual abilities, which qualify them to be philosophers, with a powerful imagination. The intellectual faculty of the philosophers and the prophets receives an overflow from the active intellect. In the case of the prophets, this overflow not only brings about intellectual activity but also passes over into the imaginative faculty, giving rise to visions and dreams. The fact that prophets have a strong imagination gives them no superiority in knowledge over philosophers, who do not have it. Moses, who belonged to a higher category than did the other prophets, did not have recourse to imagination.

The laws and religion as instituted by Moses are intended not only to ensure the bodily welfare and safety of the members of the community but also to facilitate the attainment of intellectual truths by individuals gifted enough to uncover the various hints embodied in religious laws and practices. This does not mean that all the beliefs inculcated by Judaism are true. Some indeed express philosophical truths, although in an inaccurate way, in a language suited to the intellectual capacity of the common people, who in general cannot grasp the import of the dogmas they are required to profess. Other beliefs, however, are false but necessary for the preservation of a public order upholding justice—*e.g.*, the belief that God is angry with wrongdoers.

As far as the Law—that is, the religious commandments—is concerned, two aspects of Maimonides' position may be distinguished. First, he maintained that it is unique in its excellence and valid for all time. This profession of faith, at least with regard to its assumptions about the future, lacked philosophical justification: however, it could be regarded as necessary for the survival of Judaism. Second, he asserted that certain precepts of the Mosaic Law were related to specific historical situations and the need to avoid too sharp a break with popular customs and practices, for instance, the commandments concerning sacrifice.

Influence of the *Guide of the Perplexed*

For at least four or five centuries the *Guide of the Perplexed* exercised a very strong influence in the European centres of Jewish thought; in the 13th century, when the *Guide* was twice translated into Hebrew, these centres were Spain, the south of France, and Italy. Rather paradoxically, in view of the unsystematic character of Maimonides' exposition, it was used as a standard textbook of philosophy and condemned as such when the teaching of philosophy came under attack. The performance of this function by the *Guide* was rendered possible, or at least facilitated by, the fact that from the 13th century onward the history of Jewish philosophy in European countries acquired a continuity it had never had before. This development seems to have resulted from the substitution of Hebrew for Arabic as the language of philosophical exposition. Because of the existence of a common and relatively homogeneous philosophical background—the Hebrew texts were much less numerous and less diverse than Arabic philosophical works—and the fact that Jewish philosophers reading and writing in Hebrew read the works of their contemporaries and immediate predecessors, something like a dialogue can be discerned. In striking contrast to the immediately preceding period, European Jewish philosophers in the 13th century and after frequently devoted a very considerable part of their treatises to discussions of the opinions of other Jewish philosophers. That many of the Jewish philosophers in question wrote commentaries on the *Guide* undoubtedly furthered this tendency.

Influence of Averroës and Christian Scholastics on medieval Jewish philosophy

Averroists. The influence of Maimonides' great Islamic contemporary Averroës, many of whose commentaries and treatises were translated into Hebrew, was second only to that of Maimonides on Jewish intellectual development. Indeed, it may be argued that for philosophers, as distinct from the general reading public, it often came first. In certain cases commentators on the *Guide* tend, in spite of the frequent divergences between the two philosophers, to quote Averroës' opinions in order to clarify those of Maimonides.

The apparently significant influence of Christian Scholastic thought on Jewish philosophy was often not openly acknowledged by Jewish thinkers in the period beginning with the 13th century. Samuel ibn Tibbon, one of the translators of the *Guide* into Hebrew and a philosopher in his own right, remarked on the fact that the philosophical sciences were more widely known among Christians than among Muslims. Somewhat later, at the end of the 13th century and after, Jewish scholars in Italy translated into Hebrew varied texts of Thomas Aquinas and other Christian Scholastics; not infrequently, some of them acknowledged the debt they owed their Christian masters. In Spain and in the south of France, a different convention seems to have prevailed up to the second half of the 15th century. Whereas Jewish philosophers of these countries felt no reluctance about referring to Greek, Arabic, and other Jewish philosophers, as a rule they refrained from citing Christian thinkers whose views had, in all probability, influenced them. In the case of certain Jewish thinkers, this absence of reference to the Christian Scholastics served to disguise the fact that in many essentials they were representative of the philosophical trends, such as Latin Averroism, that were current among the Christian Scholastics of their time.

Quite evident is the resemblance between certain views professed by the Latin Averroists and the parallel opinions of Isaac Albalag, a Jewish philosopher who lived in the second half of the 13th century, probably in Catalonia, Spain, and who wrote a commentary in Hebrew on the *Tahāfut al-falāsifah* ("The Inconsistencies of the Philosophers"), an exposition of Avicenna's doctrine written by the Muslim philosopher al-Ghazālī. Albalag's assertion that both the teachings of the Bible and the truths demonstrated by reason must be believed even if they are contradictory clearly poses the question whether some historical connections exist between this view and the Latin Averroist doctrine that there are two sets of truths—the religious and the philosophical—and that these are not necessarily in accord. In most other points Albalag was a follower of the system of Averroës himself. This philosophical position may be exemplified by his rejection of the view that the world was created in time. He professed, it is true, to believe in what he called "absolute creation in time." This expression, however, merely signifies that at any given moment the continued existence of the world depends on God's existence, an opinion that is essentially in harmony with Averroës.

Joseph Caspi, a prolific 14th-century philosopher and exegetical commentator, maintained a somewhat unsystematic philosophical position that seems to have been influenced by Averroës. He expressed the opinion that knowledge of the future, including that possessed by God himself, is of a probabilistic nature. The prescience of the prophets is of the same nature. It is more than likely that Caspi's interest in this problem had some connection with the debate about future contingencies in which Christian Scholastics were engaged at that time.

Moses Narboni

Moses of Narbonne, or Moses Narboni, who lived in the south of France in the 14th century, was, like many other Jewish writers of this period, mainly a writer of commentaries. He wrote commentaries on biblical books, on treatises of Averroës, and on Maimonides' *Guide*. In his commentary on the *Guide*, Narboni often interprets the earlier Jewish philosopher's opinions by recourse to Averroës' views. Narboni also expounded and gave radical interpretations to certain conceptions that he understood as implied in the *Guide*. According to Narboni, God participates in all things, because he is the measure of all substances. God's existence appears to be bound up with that of the world, to which he has a relation analogous to that existing between a soul and its body (a comparison already made in the *Guide*).

Gersonides. Gersonides (Levi ben Gershom), another 14th-century Jewish philosopher born in the south of France, wrote the systematic philosophical work *Sefer milḥamot Adonai* ("The Book of the Wars of the Lord") as well as many philosophical commentaries. Gersonides apparently never explicitly mentioned Christian Scholastic philosophers; he cited Greek, Arabic, and Jewish thinkers only, and in many ways his system appears to have stemmed from the doctrines of Maimonides or Averroës, regardless of whether he agreed with them or not. For example, he explicitly rejected Maimonides' doctrine of negative theology. A comparison of his opinions and of the particular problems that engaged his attention, with the views and debates found in Scholastic writings of his period, however, suggests that he was also influenced by the Latins on certain points.

Gersonides disagreed both with the Aristotelian philosophers who maintained the eternity of the world and with the religious partisans who believed in the creation of the world in time out of nothing. He maintained that God created the world in time out of a preexistent body lacking all form. As conceived by Gersonides, this body seems to be similar to primal matter.

The problem of human freedom of action and a particular version of the problem of God's knowledge of future contingencies form an important part of Gersonides' doctrine. Gersonides, who, unlike the great Jewish and Muslim Aristotelians, believed in astrology, held that all happenings in the world except human actions are governed by a strict determinism. God's knowledge does not extend to individual human acts but embraces the general order of things; it grasps the laws of universal determinism but is incapable of apprehending events resulting from man's freedom. Thus, the object of God's knowledge is an ideal world order, which differs from the real world insofar as the latter is in some measure formed according to man's free will.

In political and social doctrine there is a fundamental difference between Maimonides and Gersonides. Gersonides does not appear to have assigned to the prophets any political function; according to him, their role consists in the prediction of future events. The providence exercised by the heavenly bodies ensures the existence in a given political society of men having an aptitude for and exercising the handicrafts and professions necessary for the survival of the community. He remarked that in this way the various human activities are distributed in a manner superior to that outlined in Plato's *Republic*. Thus, he rejected explicitly Plato's political philosophy, which, having been adapted to a society ruled through the laws promulgated by a prophet (Muḥammad), had been an important element of Jewish philosophy in the Arabic period.

Ḥasdai Crescas. Ḥasdai ben Abraham Crescas (1340–1410), a Spanish-Jewish thinker, like Gersonides had thorough knowledge of Jewish philosophy and partial knowledge of Islāmic philosophy, and both seem to have been influenced by Christian Scholastic thought; moreover, in certain important respects Crescas was influenced by Gersonides himself. In Crescas' main work, *Or Adonai* ("The Light of the Lord"), however, one of his objectives, quite contrary to Gersonides, was to expose the weakness and insufficiency of Aristotelian philosophy. This attitude may be placed in the wider context of the return to religion itself, as opposed to the Aristotelian rationalization of religion, and the vogue of Kabbala (Jewish theosophical mysticism), both of which were characteristic features of Spanish Jewry in Crescas' time. This change in attitude has been regarded as a reaction to the increasing precariousness of the position of the Jewish community in Spain.

Anti-Aristotelianism

The low estimation of the certainties and the rationalistic arrogance of the medieval Aristotelians coincided chronologically with a certain disintegration of and disaffection toward classical Aristotelian Scholasticism. Relevant to this decline were the so-called voluntarism of Duns Scotus, the Nominalism of William of Ockham and other 13th–14th-century Christian Scholastics, and the development, in the 14th century and after, of anti-Aristotelian physics at the University of Paris and elsewhere. Significantly there is a pronounced resemblance between Crescas' views and two of these trends, Scotism and the "new" physics.

Crescas accepted Gersonides' view that divine attributes cannot be negative, but unlike his predecessor he centred his explanation of the difference between the attributes of God and those of created existents on the antithesis between an infinite being and finite beings. It is through infinitude that God's essential attributes—wisdom, for instance—differ from the corresponding and otherwise similar attributes found in created beings. In Crescas', as in Spinoza's, doctrine (see below), God's attributes are also infinite in number. The central place assigned to the thesis of God's infinity in Crescas' system suggests the influence of Duns Scotus' theology, which is similarly founded upon the concept of divine infinity.

The problem of the infinite approached from an altogether different angle was one of the main themes of Crescas' critique of Maimonides' 25 propositions; these propositions, concerned mainly with Aristotelian physical doctrines, had been set forth in the *Guide* as the basis of Maimonides' proof for the existence of God. Crescas' declared purpose in criticizing and rejecting several of these propositions was to show that the traditional Aristotelian proofs (founded in the first place on physical doctrines) were not valid. In the course of his critique, Crescas attempted to disprove the Aristotelian thesis that the existence of an actual infinite is impossible. He held that space is not a limit but a tridimensional extension, that it is infinite, and that, contrary to Aristotle, the existence of a vacuum and of more worlds than one is possible. He also criticized as being impossible the thesis of the Aristotelian philosophers that there exists an infinite number of causes and effects, which have order and gradation. This thesis refers not to a temporal succession of causes and effects that have a similar ontological status but to a vertical series, descending from God to the lowest rung in creation. His attacks were likewise directed against the Aristotelians' conceptions of time and of matter.

Crescas' fundamental opposition to Aristotelianism is perhaps most evident in his rejection of the conception of intellectual activity as the supreme state of being for man and for God. Crescas' God is not first and foremost an intellect, and the supreme goal to which man can aspire is to love God with a love corresponding as far as possible to the infinite greatness of its object and to rejoice in the observance of his commandments. God, too, loves man, and his love, in spite of the lowliness of its object, is proportionate to his infinity.

Crescas attacked the separation of the intellect from the soul as conceived by the Aristotelians and attempted, perhaps in part under the influence of Judah ha-Levi, to refute the Aristotelian doctrine that the actualized intellect, in contradistinction to the soul, survives the death of the body. According to Crescas, the soul is a substance in its own right and can be separated from the body; it continues to subsist after the body's death.

Joseph Albo. Whereas Crescas unmistakably regarded the Aristotelian philosophers as adversaries to be criticized or combatted, the attitude of Joseph Albo (*c.* 1380–*c.* 1444), who regarded Crescas as his teacher, is much less clearly defined. Albo did not eschew self-contradiction, apparently considering it a legitimate precaution on the part of a philosophical or theological author; indeed, he

Albo's mixture of religious tradition and rational philosophy

indulged in it in a much more obvious way than did Maimonides. But, whereas the latter's fundamental philosophical position is fairly clear, the problem being how far he was prepared to deviate from Aristotelian doctrine in the interests of religion, there may be valid doubt whether Crescas and the Jewish religious tradition or Maimonides and Averroës were Albo's true masters. Mainly because of this perhaps deliberate failure to explain to the reader where he really stood, Albo has often been dismissed as an eclectic. He was strongly influenced not only by the authors just mentioned but also by Sa'adia. He seems to have had a considerable knowledge of Christian theology, even adopting for his own purposes certain Scholastic doctrines. He differs from Crescas and to some extent resembles Maimonides in having a marked interest in political theory.

The proclaimed theme of Albo's magnum opus, *Sefer ha-'iqqarim* ("Book of Principles"), is the investigation of the theory of Jewish religious dogmas, the number of which Maimonides, in a nonphilosophical work, had set at 13, whereas Albo, following a doctrine that in the last analysis seems to go back to Averroës, would limit the number to three: the existence of God, divine providence in reward and punishment, and the Torah as divine revelation. One section, usually including the philosophical and the traditional religious interpretations side by side, is devoted to each of these dogmas. Albo's principal and relatively novel contribution to Jewish doctrinal evolution is the classification, in his introduction, of natural, conventional, and divine law.

Natural law (the universal moral law inherent in human nature) is necessary, because man, being political by nature, must belong to a community, which may be restricted in size to one town or may extend over the whole earth. Natural law preserves society by promoting right and repressing injustice; thus, it restrains men from stealing, robbing, and murdering. The positive laws instituted by wise men take into account the particular nature of the people for whose benefit they are instituted, as well as other circumstances. This means that they differ from the natural law in not being universally applicable. Neither natural law nor the more elaborate conventional laws, however, lead men toward true spiritual happiness; this is the function of divine laws instituted by a prophet, which teach men true theoretical opinions. Whereas Maimonides maintained that Judaism was the only divine law promulgated by a true prophet, Albo considered that the commandments given to Noah for all mankind also constitute divine law, which ensures, although to a lesser degree than does Judaism, the happiness of its adherents. This position justifies a certain universalism; in accordance with a Talmudic saying, Albo believed that the pious among the non-Jews—that is, those who observe Noah's laws—have a share in the world to come. But he rejected the pretensions of Christianity and Islām to be divine laws.

MODERN PHILOSOPHY

The Iberian-Dutch philosophers. The expulsion of the Jews from Spain and Portugal produced a new centre of Jewish thought, Holland, where many of the exiled Jews found a new and safer domicile; the tolerance of the regime seemed to provide guarantees against external persecution. This did not prevent, and indeed may have furthered, the establishment of an oppressive internal orthodoxy that was prepared to chastise rebellious members of the community. This was evident in the cases of Uriel Acosta, or da Costa, and Benedict de Spinoza, two 17th-century philosophers who rebelled against Jewish orthodoxy and who were excommunicated for their views (Acosta twice).

The two excommunicados: Acosta and Spinoza

Uriel Acosta. Acosta came to Amsterdam from Portugal, where, belonging to a family of Marranós (Jews who had converted to escape religious persecution), he had been brought up in the Catholic faith; his philosophical position was to a great extent determined by his antagonism to the orthodox Judaism that he encountered in Amsterdam. His growing estrangement from generally accepted Jewish doctrine is attested by his Portuguese treatise *Sobre a Mortalidade da Alma* ("On the Mortality of the Soul"). He considered that the belief in the immortality of the soul has had many evil effects, for it impels men to choose an ascetic way of life and even to seek death. According to him, nothing has tormented men more than the belief in an inner, spiritual good and evil. At this stage Acosta affirmed the authority of the Bible from which, according to him, the mortality of the soul can be proved.

In his autobiography, written in Latin and entitled *Exemplar Humanae Vitae* ("Example of a Human Life"), he takes a more radical position. He proclaims the supreme excellency of the natural moral law (which, when arguing before Jews, he seems to identify with the divine commandments to Noah, thus suggesting a correspondence with the view of Albo). Accordingly, he denies the validity of the argument that natural law is inferior to Judaism and Christianity, because he believes that both these religions teach the love of one's enemies, a precept that is not a part of natural law and is a manifest impossibility.

Benedict de Spinoza. Although modern philosophers of Jewish origin are not considered as belonging to the history of Jewish philosophy unless they deal with Judaic themes, this restriction may not apply to Spinoza for the following reasons. (1) It was through the study of Jewish philosophical texts that Spinoza was first initiated into philosophy. (2) Spinoza's system is in part a radicalization of, or perhaps a logical corollary to, medieval Jewish doctrines; and the impact of Maimonides and of Crescas is evident. (3) A considerable portion of Spinoza's *Tractatus Theologico-Politicus* deals with problems related to Judaism.

Spinoza's view of prophecy

The first chapters of the *Tractatus* show that the doctrine of prophecy is of central importance to Spinoza's explanation of Judaism and that, in dealing with this subject, he used Maimonides' categories, although he applied them to different people or groups of people. Maimonides held that the prophets combined intellectual perfection, which made them philosophers, with perfection of the imaginative faculty. He also referred to a category of persons endowed with a strong imagination but possessing no extraordinary intellectual gifts; this category includes, for example, lawgivers and statesmen. Spinoza took over this last category but applied it to the prophets, whom he described as possessing vivid imaginations but as not necessarily having outstanding intellectual capacities. He denied that the biblical prophets were philosophers and used a philosophical and historical approach to the Scriptures to show that the contrary assertion is not borne out by the texts.

Spinoza also denied Maimonides' assertion that the prophecy of Moses was essentially different from that of the other prophets and that this was largely because Moses, in prophesying, had no recourse to the imaginative faculty. According to Spinoza, the distinctive fact about Moses' prophecy was that he heard the voice of God in a prophetic vision—that is, in a state in which his imagination was active. In this assertion Spinoza employed one of Maimonides' categories of prophecy, differentiated in the *Guide* according to certain characteristics of prophetic dreams and visions; however, Maimonides thought it improbable that the voice of God was ever heard in prophetic vision, and he held that this category is purely hypothetical. It seems evident that in his classification of Moses, Spinoza was concerned not with what really happened in history but with pigeonholing the biblical evidence into Maimonides' theoretical framework so that it fit in with his own theologico-political purpose: to show that there could be a religion superior to Judaism.

This purpose made it imperative to propound in the *Tractatus Theologico-Politicus* a theory concerning Jesus, whom Spinoza designates as Christus. The category and the status assigned to Jesus are by and large similar to those that Maimonides attributed to Moses. Thus, Jesus is referred to in the *Tractatus* as a religious teacher who makes recourse not to the imaginative faculty but solely to the intellect. His authority may be used to institute and strengthen the religion Spinoza called *religio catholica* ("universal religion"), which has little or nothing in common with any of the major manifestations of historic Christianity.

The difference between Judaism and Spinoza's *religio* *catholica* corresponds to the difference between Moses and Jesus. After leaving Egypt the Jews found themselves, in Spinoza's view, in the position of people who had no allegiance to any positive law; they had, as it were, reverted to a state of nature and were faced with the need to enter into a social pact. They were also an ignorant people and very prone to superstition. Moses, a man of outstanding ability, made use of the situation and the characteristics of the people in order to make them accept a social pact and a state founded upon it that, contrary to Spinoza's scheme for his ideal communities, were not based first and foremost upon utilitarian—that is, reasonable—consideration of the advantages of life in society over the state of nature.

The social pact concluded by the children of Israel in the desert was based upon a superstitious view of God as "King" and "Judge," to whom the children of Israel owed whatever political and military successes they obtained. It was to God rather than to the representatives of the popular will that the children of Israel transferred political sovereignty. In due course political sovereignty was vested in Moses, God's representative, and in his successors. It should be added that, in spite of Spinoza's insistence on the superstitious foundations of the ancient Israelite state, his account of its regime was not wholly unsympathetic. He believed, however, that it contained the seeds of its own destruction and that, with the extinction of this state, the social pact devised by Moses had lapsed and all the political and religious obligations incumbent upon the Jews had become null and void.

The rational state and its religion

It could be argued that, because the state conceived by Spinoza is based not on superstitious faith but on a social contract originating in rational, utilitarian considerations, it does not necessarily need to have its authority safeguarded and stabilized by means of religion. Nevertheless, Spinoza appears to have held the view—perhaps derived from a purely empirical knowledge of the behaviour of the common run of men—that there is a need for religion. In order to fulfill the need for some religion and to obviate the danger of harmful religions, he devised his *religio catholica,* the universal religion, which has the following distinctive traits: (1) Its main purpose, a practical one (which is furthered by recourse to the authority of Jesus), is to impel men to act in accordance with justice and charity. Such conduct is tantamount to obedience to the laws of the state and to the orders of the magistrates, in whom sovereignty is vested; for disobedience—even if it springs from compassionate motives—weakens the social pact, which safeguards the welfare of all the members of the community; in consequence, its evil effects outweigh whatever good it may produce. (2) Although religion, according to Spinoza, is not concerned with theoretical truth, in order to be effective the *religio catholica* requires dogmas, which he set forth in the *Tractatus.* These dogmas are formulated there in terms that can be interpreted in accordance both with the philosophical conception of God that Spinoza regarded as true and also with the superstitious ideas of ordinary people. It follows that if they are accepted as constituting by themselves the only creed that everybody is obliged to profess, people cannot be persecuted on account of their beliefs; Spinoza held that such a persecution may lead to civil war and may thus destroy the state. Philosophers are free to engage in the pursuit of truth and to attain, if they can, the supreme goal of man, freedom grounded in knowledge. There can be little doubt that the furtherance of the cause of tolerance for philosophical opinions was one of Spinoza's main objects in writing the *Tractatus.*

The relation between Spinoza's *Ethics,* his major philosophical work, and Jewish medieval philosophy is much more ambiguous than in the case of Spinoza's *Tractatus Theologico-Philosophicus.* In a way, Spinoza's metaphysical system, contained in the *Ethics,* can be regarded as being, in part, a spelling out of some extreme consequences, which could perhaps be legitimately drawn from medieval Aristotelianism; but the facts of the case are no doubt much more complicated than this.

German philosophers. *Moses Mendelssohn.* Moses Mendelssohn opened what may be called the German period of Jewish philosophy (*c.* 1750–*c.* 1830). This period, in which a considerable number of works on Jewish philosophy were written in German and often under the influence of German philosophy, is also marked by the emancipation of the Jews—that is, by the abrogation of discriminatory laws directed against them—and by their partial or complete assimilation. In this period in particular, the term Jewish philosophy applies especially to works the main purpose of which, or one of the main purposes of which, consists in proposing a definition of Judaism and a justification of its existence. The second task is often conceived as necessitating a confrontation of Judaism with Christianity rather than with philosophy, which served as a critical point of comparison for many medieval philosophers. This change seems to have been a result of the demarcation of the sphere of religion in such a way that, at least in the opinion of the philosophers, possible points of collision no longer existed between it and philosophy. This demarcation was largely furthered by the doctrine of Spinoza, from whom Mendelssohn and others took over and adopted for their own purposes certain fundamental ideas concerning Judaism. Like Spinoza, Mendelssohn held that it is not the task of Judaism to teach rational truths, although they may be referred to in the Bible. Contrary to what he called Athanasian Christianity—that is, the doctrine set forth in the Athanasian Creed—Judaism has no binding dogmas; it is centred on inculcating belief in certain historical events and on action—that is, observance of religious law (including the ceremonial commandments). Such observance is supposed to lead to happiness in this world and in the afterlife. Mendelssohn did not reject this view offhand, as Spinoza would have done; indeed, he seems to have been prepared to accept it, God's mysteries being inscrutable, and the radicalism and what may be called the consistency of Spinoza being the complete antithesis of Mendelssohn's apologetics. Non-Jews were supposed by Mendelssohn to owe allegiance to the natural moral law.

Solomon Formstecher. Whereas Mendelssohn continued the medieval tradition, at least to some extent, or adopted Spinoza's doctrine for his purposes, the Jewish philosophers of the first half of the 19th century may generally be regarded as disciples of the philosophers of their own time. In *Die Religion des Geistes* ("The Religion of the Spirit"), Solomon Formstecher (1808–89) may have been influenced by F.W.J. Schelling, an eminent German philosopher, in his conception of nature and spirit as manifestations of the divine. There are, in Formstecher's view, two types of religions that correspond to these manifestations: (1) the religion of nature, in which God is conceived as the principle of nature or as the world soul, and (2) the religion of the spirit, which conceives of God as an ethical being. According to the religion of the spirit, God has produced the world as his manifestation in full freedom and not, as the religion of nature tends to profess, because the world was necessary for his own existence.

The religions of nature and of spirit

The religion of the spirit, which corresponds to absolute religious truth, was first manifested in the Jewish people. The religious history of the world may be understood as a process of universalization of the Jewish religion. Thus, Christianity propagated Jewish conceptions among the nations; however, it combined them with pagan ideas. The pagan element is gradually being eliminated—Protestantism, for instance, in this respect, marks considerable progress. When at long last the Jewish element in Christianity is victorious, the Jews will be right to give up their isolation. The progress that will bring about this final religious union is already under way.

Samuel Hirsch. The main philosophical work of Samuel Hirsch, entitled *Die Religionsphilosophie der Juden* ("The Philosophy of Religion of the Jews"), was decisively influenced by G.W.F. Hegel. This influence is most evident in Hirsch's method and in the task that he assigned to the philosophy of religion—the transformation of religious consciousness into conceptual truth. Contrary to Hegel, however, he did not consider religious truth to be inadequate as compared to philosophical truth.

God revealed himself in the first stages of Jewish history by means of miracles and of prophecy. At present, he manifests himself in the miracle that is constituted by the existence of the Jewish people. At its beginning in the time of Jesus, Christianity was identical with Judaism. The decisive break between the two religions was caused by Paul. When the Pauline elements are eliminated from Christianity, it will be in all essentials in agreement with Judaism, which, however, will preserve its separate existence.

Nachman Krochmal. Nachman Krochmal, a native of Galicia (at that time, part of Austria), was the author of *More nevukhe ha-zman* ("Guide for the Perplexed for Our Time"), a treatise in Hebrew on the philosophy of history and on Jewish history that had a considerable influence. Krochmal's philosophical thought was centred on the notion of "spirit," Krochmal being mainly concerned with the "national spirit," the particular spirit that is proper to each people and that accounts for the peculiar characteristics differentiating one people from another in every domain of human activity. The national spirits of all peoples except the Jewish are, according to Krochmal, essentially particular. Hence, the national spirit either becomes extinct with the extinction of the nation or, if it is a powerful spirit, is assimilated by some other nation. The Jewish people have a special relation to the Universal Spirit, who is the God of Israel. This relation accounts for the perpetuity of the Jewish people.

Solomon Steinheim. Solomon Ludwig Steinheim, the author of *Die Offenbarung nach dem Lehrbegriff der Synagoge* ("The Revelation According to the Doctrine of the Synagogue"), was apparently influenced by the antirationalism of Friedrich Heinrich Jacobi, a German philosopher. His criticism of science is based on Jacobi's criticism, but he did not agree with Jacobi in opposing discursive reason to the intuitive knowledge of God—Steinheim contrasted human reason to divine revelation. The main point on which the revelation, vouchsafed to the prophets of Israel, is opposed to reason is to be found in the fact that the God posited by reason is subject to necessity, that he can act only in accordance with laws. Moreover, reason affirms that nothing can come from nothing. Accordingly, God is free to create not a good world, but only the best possible world. Revealed religion, on the other hand, affirms the freedom of God and the creation of the world out of nothing.

The Neo-Kantian philosopher of Judaism

Hermann Cohen. There seems to be little connection between the Jewish philosophers of the first half or two-thirds of the 19th century and Hermann Cohen (1842–1918), the head of the Neo-Kantian school centred at the University of Marburg. In a certain sense Cohen may be regarded as a rather unusual case among the philosophers of Judaism of his and the preceding generations because of the two aspects of his philosophical thought—the general and the Jewish—and the uneasy equilibrium between them. Judaism was by no means the only important theme of his philosophical system; it was one of several and not even his point of departure. There is no doubt that, for most of his life, Cohen was wholly committed to his brand of Kantianism, in the elaboration of which he displayed considerable originality—it has been maintained with some justification that his doctrine manifests a certain (unintentional) kinship with Hegel's. Cohen's idea of God, however, derives from an analysis and a development of certain conceptions of Immanuel Kant. In Cohen's view reason requires that nature be conceived of as conforming to one rational plan and that harmony exist between the domains of natural and of moral teleology (ultimate purposes or ends). These two requirements, in turn, necessitate the adoption of the idea of God—the word idea being used in the Kantian sense, which means that no assertion is made about the metaphysical reality of God.

Cohen seems to have changed his attitude in the last years of his life; although he did not explicitly renounce his previous positions, a considerable shift of emphasis can be discerned in his doctrines. The notion of the human individual—an individual who is weak and full of sin—comes to the fore, as well as the conception of a correlation, a relationship, between God and the individual. This relationship is one of love, the love of God for man and the love of man for God. It is difficult to reconcile the conception of God expounded in Cohen's work of his last period with his Kantian or Neo-Kantian attitude toward metaphysics.

Franz Rosenzweig. Franz Rosenzweig published his main philosophical work, *Der Stern der Erlösung* (*The Star of Redemption,* 1971), in 1921. This work begins with a rejection of the traditional philosophical attitude that denies the fear of death, maintaining, instead, that this fear is the beginning of the cognition of the All. Man should continue to fear death, despite the indifference of philosophy and its predilection for accepting death. Traditional philosophy is interested exclusively in the universal, and it is monistic—its aim is to discover one principle from which everything can be derived. This tendency of philosophy, however, denatures human experience, which knows not one but three separate domains (which Kant had referred to in a different context), namely, God, the world, and man.

According to Rosenzweig, God (like the world and like man) is known through experience (the experience of revelation). In Greek paganism, the most perfect manifestation of paganism, every one of these domains subsists by itself: the gods, the cosmos, and man as the tragic, solitary, silent hero. Biblical religion is concerned with the relation between the three: the relation between God and the world, which is creation; the relation between God and man, which is revelation; and the relation between man and the world, which leads to salvation. The philosophy that renounces the ambition to find one principle for everything that exists and that follows biblical religion in centring on the connections between the three domains and between the words and acts that bring about and develop these connections, Rosenzweig termed the narrative philosophy; the term and the concept were taken over from Schelling, whose influence Rosenzweig repeatedly emphasized.

The biblical faith brought forth two valid religions—Christianity and Judaism. The first is described by Rosenzweig as the eternal way; the Christian peoples seek in the vicissitudes of time and history the way to salvation. In contradistinction to them, the existence of the stateless Jewish people is not concerned with time and history; it is—notwithstanding the hope for final salvation—already an eternal life, renewed again and again according to the rhythm of the liturgical Jewish year.

Martin Buber. Since the early years of the 20th century, Martin Buber has exercised a powerful influence on both Jews and non-Jews. In his early period Buber was led, partly through empathy with Jewish and non-Jewish mysticism, to stress unitive experience and knowledge, in which the difference between one man and another and between man and God tend to disappear. But in his final period he taught, following, as he claimed, a suggestion of Ludwig Feuerbach, a 19th-century German philosopher, that man can only realize himself as a human being in a relation with another, who may be another man or God. This conception of the "I and Thou" relationship leads to the formulation of Buber's view of the dialogical life—the mutual, responsive relation between man and others—and accounts for the importance that he attaches to the category of "encounter." (S.Pi.)

"I and Thou"

Jewish mysticism

This section deals with the special nature and characteristics of Jewish mysticism, the main lines of its development, and its role in present-day religion and culture.

NATURE AND CHARACTERISTICS

The Judaic context. The term mysticism applies whenever a person is convinced that it is possible to establish *direct contact,* apart from sense perception and intellectual apprehension, with the divine—a reality undefinable by pure logic and believed to be the ultimate ground of being. Since mysticism springs from an aspiration to join and grasp that which falls outside ordinary experience, it is not easily restricted within precise limits. The boundary line that separates mysticism from metaphysics and cosmology (doctrines on the basic nature or structure of being and the world), from theosophies (systems of thought claiming special insights or revelation into the divine nature), and various forms of occultism (the study and control of supernatural powers), and from theurgy (the art of compelling or persuading divine powers) and even magic, often of the lowest kind, is not clear.

If mysticism is defined as the search for *direct* contact with the divine, however, it seems to be incompatible with Judaism. In its classical and normative form, Judaism appears as faith in a sole God who created the universe and who chose to reveal himself to a selected group by means of a rule of life he imposed on it—Torah ("Guidance" or "Teachings," incorrectly rendered as "Law"). The earthly destiny of the chosen nation, as well as the eternal salvation of the individual, in traditional Judaic beliefs, depends upon the observance of this rule of life, through which any relationship to God must take place. The fact is, however, that in the religious history of Judaism the quest for God goes beyond this relationship mediated by Torah, without ever dispensing with it (since that would take the seeker outside of Judaism) or pretending to reach the depths of the mystery of the divine, or still less to end in an ontological identification with it (where God and man are the same in nature and being).

Three types of Jewish mysticism. Three types of mysticism may be discerned in the history of Judaism: the ecstatic, the contemplative, and the esoteric. Though they are distinct types, in practice there are frequent overlappings and mixtures between them.

The first type is characterized by the quest for God—or, more precisely, for access to a supernatural realm, which is itself still infinitely remote from the inaccessible deity—by means of ecstatic experiences; this method is sometimes tainted by theurgy. The second follows the way of metaphysical meditation pushed to the limit, always bearing in its formulations the imprint of the cultural surroundings of the respective thinkers, who are exposed to influences from outside Judaism; this was the case with Philo of Alexandria (*c.* 15 BCE–after 40 CE) and a few of the Jewish thinkers of the Middle Ages, who drew their inspiration from Greco-Arabic Neoplatonism and sometimes also from Muslim mysticism.

Esoterism

The third type of mysticism claims an esoteric knowledge (hereafter called esoterism) that explores the divine life itself and its relationship to the extradivine level (the natural, finite realm) of being, a relationship that is subject to the "law of correspondences." From this perspective, the extradivine is a symbol of the divine; that is, a reality that reveals another, superior reality, whence reciprocal action of the one on the other (which corresponds to it) exists. This form of mysticism, akin to gnosis—the secret knowledge claimed by Gnosticism, a Hellenistic religio-philosophical movement—but purged, or almost purged, of the dualism that characterizes the latter, is what is commonly known as Kabbala (literally "tradition"). By extension, this term is also used to designate technical methods, used for highly diverse ends, ranging from the conditioning of the aspirant to ecstatic experiences to magical manipulations of a frankly superstitious character. If the concept of spiritual energy acting on matter and at a distance originally underlay these practices, it finally became unrecognizable and all that remained was a collection of "tricks of the trade."

The favour with which the doctrine of correspondences was regarded by ancient and medieval science, as well as the tendency in the three monotheistic religions (Judaism, Christianity, and Islām) to reconcile the results of rational reflection with the data of revelation, had the result of turning speculation on the origin and order of the universe toward mysticism.

Perennial human questions

It must also be noted that the quest for God implies the search for solutions to problems that go beyond those of religion in the narrow sense and that arise even when there is no interest in the relationship between man and supernatural powers. Man ponders the problems of his origins, his destiny, his happiness, his suffering—questions that arise outside of religion, as well as within nonmystical forms of religious life; the presence or absence of religious institutions or dogmas is of little importance when it comes to these questions. They were all formulated within nonmystical Judaism and served as the basis and framework for the setting and solution of problems in the various forms of Jewish mysticism. This mysticism, especially in its "Kabbalistic" form, brought about profound transformations in the concepts of the world, God and "last things" (resurrection, last judgment, messianic kingdom, etc.) set forth in biblical and rabbinical Judaism. Nevertheless, Jewish mysticism's own set of problems about the origins of the universe and of man, of evil and sin, of the meaning of history, of the afterlife and the end of time is rooted in the very ground of Judaism and cannot be conceived outside of an exegesis of revealed Scripture and rabbinical tradition.

MAIN LINES OF DEVELOPMENT

A study of the main lines of Jewish mysticism, following its actual historical development, reveals that during a very long period, from its origins in the 1st century CE to the middle of the 12th century, only the first two of the three types outlined above existed. It was not until the second half of the 12th century that esoterism became clearly discernible; from then on it continued to develop in various forms up to very recent times.

Early stages to the 6th century CE. The centuries that followed the return from the Babylonian Exile in the 6th century BCE witnessed the growth and intensification of reflection on the intermediary beings between man and God, of meditation on the divine appearances whose special place of occurrence had formerly been the most sacred part of the Jerusalem Temple, of speculation on the coming into being and organization of the universe and on the creation of man. None of these themes was absent from the Bible, which was held to be divinely revealed, but each had become the object of a constant ideological readjustment that also involved the infiltration of concepts from outside and reaction against them. The speculative taste of Jewish thinkers between the 2nd century BCE and the 1st century CE took them in many different directions: angelology (doctrine about angels) and its counterpart demonology (doctrine about devils); mythical geography and uranography, description of the heavens; speculation on the divine manifestations—which had as background the Jerusalem Temple worship and the visions of the moving "Throne" (the "Chariot," *Merkava*) in the prophecy of Ezekiel; on the double origin of man, a being formed of the earth but also the "image of God"; on the end of time; on resurrection (a concept that appeared only toward the end of the biblical period); and on rewards and punishments in the afterlife.

The literary crystallization of all this ferment was accomplished in writings, such as the book of Enoch, of which Pharisaic (rabbinical) Judaism—which became the normative Jewish tradition after the Roman conquest of Jerusalem and the destruction of the Second Temple (70 CE)—retained almost nothing and even the vestiges of which it tended to obliterate in its own writings; the Talmud and the Midrash (rabbinical legal and interpretative literature) touched these themes only with great reserve, often unwillingly and more often in a spirit of negative polemic.

As early as the 1st century CE, and probably even before the national calamity of 70, there were certainly sages or teachers recognized by the religious community for whom meditation on the Scriptures—especially the creation narrative, the public revelation of the Torah on Mount Sinai, the *Merkava* vision of Ezekiel, the Song of Solomon—and reflection on the end of time, resurrection, and the afterlife were not only a matter of exegesis and of attaching new ideas to texts recognized to be of divine origin but also a matter of inner experience. It was, however, probably in other circles that speculation on the invisible world was engaged in and where the search for the means of penetrating it was carried out. It is undeniable that there exists a certain continuity between the apocalyptic visions (*i.e.*, of the cataclysmic advent of God's Kingdom) and documents of certain sects (Dead Sea Scrolls) and the writings, preserved in Hebrew, of the "explorers of the supernatural world" (*Yorde Merkava*). The latter comprise ecstatic hymns, descriptions of the "dwellings" (*hekhalot*) located between the visible world and the ever-inaccessible divinity, whose transcendence is paradoxically expressed by anthropomorphic descriptions consisting of inordinate hyperboles (*Shi'ur qoma*, "Divine Dimensions"). In addition, a few documents have been preserved that attest to the existence of methods and practices having to do with the initiation of carefully chosen persons who were made to undergo tests and ordeals in accordance with psychosomatic criteria borrowed from physiognomy (art of determining character from physical, especially facial, traits). Some theurgic efficacy was attributed to these practices, and there was some contamination from Egyptian, Hellenistic, or Mesopotamian magic. (A curious document in this respect, rich in pagan material, is the *Sefer ha-razim*, the "Treatise on Mysteries," which was discovered in 1963.)

Two early sources of mystic writings

In this extrarational domain, there are many similarities between concepts reflected in unquestionably Jewish texts and the documents of contemporary non-Jewish esoterism, to the point that it becomes difficult, sometimes impossible, to distinguish the giver from the receiver. Two facts are certain however. On the one hand Gnosticism never ceases to exploit in its own way biblical themes (such as the tale of creation and speculation on angels and demons) that have passed through Judaism, whatever their original source may have been; on the other hand, though Jewish esoterism may borrow this or that motif from ancient gnosis or syncretism (fusion of various faiths) and may even raise to a very high rank in the hierarchy of being a supernatural entity such as the angel Metatron, also known as "little Adonai" (*i.e.*, little Lord or God), it still remains inflexibly monotheistic and rejects the Gnostic concept of a bad or simply inferior demiurge who is responsible for the creation and governing of the visible world. Finally, it is noteworthy that during the centuries that separate the Talmudic period (2nd to 5th centuries AD) from the full resurgence of Jewish esoterism in the middle of the 12th century, the texts that have been preserved progressively lose their density and affective authenticity and become reduced to the level of literary exercises that are more grandiloquent than substantial.

Sefer Yetzira. In the ancient esoteric literature of Judaism, a special place must be given to the *Sefer Yetzira* ("Book of Creation"), which deals with cosmogony and cosmology (the origin and order of the universe). Creation, it affirms with a clearly anti-Gnostic insistence, is the work of the God of Israel and took place on two different levels: the ideal, immaterial level and the concrete level. This was done according to a complex process that brings in the 10 numbers (*sefirot*, singular *sefira*) of decimal notation and the 22 letters of the Hebrew alphabet. The 10 numbers are not to be taken merely as arithmetical symbols: they are cosmological factors, the first of which is the spirit of God—with all the ambiguities that this term *ruaḥ* has in Hebrew—while the nine others seem to be the archetypes of the three elements (air, water, fire) and the spatial dimensions (up, down, and the four cardinal points). After having been manipulated either in their graphic representation or in combination, the letters of the alphabet, which are considered to be adequate transcriptions of the sounds of the language, are in turn instruments of creation.

The ontological value of speech

The basic idea of all this speculation is that speech (that is, language composed of words, which are in turn composed of letters/sounds) is not only a means of communication but also an operational agent destined to produce being—it has an ontological value. This value, however, does not extend to every form of language; it belongs to the Hebrew language alone.

The universe that is produced by means of the *sefirot* and the letters is constituted according to the law of correspondences between the astral world, the seasons that mark the rhythm of time, and man in his psychosomatic structure.

The "Book of Creation" certainly does not proceed entirely from biblical data and rabbinical reflection upon them; certain Greek influences are discernible, even in the vocabulary. What is important, however, is its influence on later Jewish thought, down to the present time: philosophers and esoterists have vied with one another in commentating it, pulling it in their own direction, and adjusting it to their respective ideologies. Even more important is the fact that Kabbala (see below *The making of the Kabbala*) borrowed a great part of its terminology from it (*sefira*, among others), naturally making semantic adaptations as required.

The speculation traced above developed during the first six centuries of the Common Era, both in Palestine and in Babylonia (later called Iraq); Babylonian Judaism had its own social and ideological characteristics, which put it in opposition to Palestinian Judaism in various aspects, including esoterism as well as other manifestations of the life of the spirit. The joint doctrinal influence of the two centres was to spread during the period from the mid-8th to 11th century among the Jews established in North Africa and Europe; mystical doctrines also filtered in, but very little is known about the circumstances and means of their penetration.

The Arabic-Islāmic influence (7th–13th century). Arabic-Islāmic culture provided another important influence in Jewish mystical development. A considerable part of Jewry, which had fallen under Muslim domination in the 7th and 8th centuries, participated in the new Arabic-Islāmic civilization; the Jews of Asia, Africa, and Spain soon adopted Arabic, the prevailing language of culture and communication. By way of Arabic-language culture, elements of Greek philosophy and Islāmic mysticism penetrated Judaism and contributed to the deepening of certain theological concepts that were Jewish in origin but had become the common property of the three religions of the Book: affirming the divine unity, purging all anthropomorphism from the idea of God, and approaching the divine by progressing on a spiritual path that leads through an ascetic discipline (both physical and intellectual) to a detachment from this world and a freeing of the soul from all that distracts it from God. Greek philosophy and Islāmic mysticism, moreover, raised very serious questions that threatened many traditional beliefs, such as the creation of the world, the providential action of God, miracles, eschatology (doctrines about the resurrection of the body, rewards, and especially material punishments in the hereafter). Even in the Christian West, where cultural contacts between the majority society and the Jewish minority were far from reaching the breadth and intensity of the Judeo-Arab relations, Jewish intellectuals were unable to remain totally impervious to the incursions of the surrounding civilization. Moreover, at the beginning of the 12th century, if not earlier, European Judaism received part of the intellectual Arabic and Judeo-Arabic heritage through translations or adaptations into Hebrew, its only cultural language.

Greek philosophy and Islāmic mysticism

The making of the Kabbala (c. 1150–1250). It was in these circumstances that, starting around 1150, manifestations of markedly theosophic ideologies appeared in the south of present-day France (in the regions of Provence–Languedoc–Roussillon). Two types can be distinguished at the outset, which are very different as to their manner of appearance, their form, and their content.

Sefer ha-bahir. The first type is represented in fragmentary, poorly written, and badly assembled texts that began to circulate in Provence–Languedoc during the third quarter of the 12th century. Their inspiration, however, leaves no doubt as to the community of their origin. They were in the form of a Midrash; that is, an interpretation of Scripture with the help of a particular interpretative method, full of sayings attributed to ancient rabbinical authorities. This whole body of texts, probably imported from the Near East (Syria–Palestine–Iraq), is known as the "Midrash of Rabbi Nehunya ben Haqana" (from the name of a 2nd-century rabbi) or *Sefer ha-bahir*, "Book of Brightness" (from a characteristic word of the first verse of Scripture to be elucidated in the work). The authorities cited are all inauthentic (as was often the case in late works), and the content of this Midrash, even its nonmystical content, is entirely Gnostic; a Gnosticism that tries nevertheless to escape any ontological dualism (and, as a matter of fact, succeeds).

Its object is to present the origin of things and the course of history centred naturally on that of the chosen people, with the vicissitudes caused in turn by obedience to God and by sin, as bound and conditioned by the

Role of the divine powers

manifestation of divine powers. These "powers" are not "attributes" derived and defined by philosophical abstraction, although that is one of the terms used to designate it: they are hypostases (essences or substances). They are inseparable from God, but each one is clothed in its own personality, each operates in its own manner, in the leaning toward severity or mercy, in dynamic correspondence with the behaviour of man, especially of the Jew, in the visible world. They are ranked in a hierarchical order, which is not yet as fixed as it became starting with the second generation of Kabbalists in Languedoc and Catalonia (see below *The school of Gerona* [*Catalonia*]). The rich nomenclature used to designate the "powers" exploits the resources of both the Bible and rabbinical tradition, of the "Book of Creation," of some ritual observances, and also of the letters of the Hebrew alphabet and the signs that can be added to them to indicate the vowels. All of this combines to give a symbolical rendering of the myth, cosmology, sacred history, and eschatology through which an anonymous group of theosophists attempt to formulate their doctrine: a Gnostic myth, except for the adjustments that eliminate the radical depreciation of the visible world.

Thus, according to the *Sefer ha-bahir*, the universe is the manifestation of the hierarchically organized divine powers, and the one that is at the bottom of the hierarchical ladder has special charge of the visible world. This entity is highly complex. Undoubtedly there are survivals of Gnostic speculation on Sophia ("Wisdom"), who is involved, sometimes to her misfortune, in the material world. This power is also the divine "Presence" (*Shekhina*) of rabbinical theology but profoundly transformed: it has become a hypostasis; by a bold innovation, moreover, it is characterized as a feminine being and thus finds itself, while remaining an aspect of the divinity, in the position of a daughter or a wife, who owns nothing herself and receives all from the father or the husband. It is also identified with the "Community of Israel," another radical innovation, but facilitated by ancient speculation based on the allegorical interpretation of the Song of Solomon, which represents the relationship of God to the chosen nation in terms of the marriage bond. Thus a theosophical equality is established between the whole of the people chosen by God, constituted into a kind of mystical body, and an aspect of the divinity, whence the solidarity and linked destiny of the latter and the human group in question. As a matter of fact, a comparable relationship between the "Presence" and Israel was not totally foreign to ancient rabbinical theology. In this light, the obedience or disobedience of Israel to its particular vocation is a determining factor of cosmic harmony or disruption and extends to the inner life of the divinity. This is the essential and definitive contribution of the *Sefer ha-bahir* to Jewish theosophy. In the same document may be seen the resurgence of a notion fought against by the older theologians—that of *metensōmatōsis*, the reincarnation into several successive bodies of a soul that has not attained the required perfection in a previous existence.

School of Isaac the Blind. Parallel to the appearance of the *Sefer ha-bahir* but independent of it, another theosophic tendency unfolded in Languedoc, the second type referred to above. The two movements would take only about thirty years to converge, to constitute what may conveniently, though not quite precisely, be called classical Kabbala. The second school flourished in Languedoc during the last quarter of the 12th century and crossed the Pyrenees into Spain in the first years of the 13th century.

The most eminent spokesman of this school was Isaac ben Abraham, known as Isaac the Blind. For this theosophist, among whose extant works there is in particular a very obscure commentary of the "Book of Creation," the general vision of the universe proceeds, to use the words of Gershom G. Scholem (the eminent 20th-century Kabbala scholar), from the link he discovers between the hierarchical orders of the created world and the roots of all beings implanted in the world of the *sefirot*. One can already see a Neoplatonic influence in the reflections of Isaac; *e.g.*, the proceeding of things from the One and the corresponding return to the heart of the primordial undifferentiatedness, which is the fullness of being and at the same time every

Return to the undifferentiated One

conceivable being. This return is not merely eschatological and cosmic but is in some way realized in the life of prayer of the contemplative mystic privileged to have supernatural inspirations, "appearances" of the prophet Elijah, by means of concentration, of orientation of action and thought (*kawwana*), and of "adhesion" (*devequt*), being-with-God, though not, indeed, a transforming union by which the human personality blends completely into the deity or becomes one with it.

The synthesis of the themes of the *Bahir* and the cosmology of the "Book of Creation," accomplished by Isaac or by others in the doctrinal environment inspired by his teachings, is and remains the foundation of Kabbala whatever enrichment, adjustments, even changes of orientation and sometimes radical modifications the composite may have undergone subsequently.

The 10 sefirot. It is also in this environment that the nomenclature of the 10 *sefirot* became more or less fixed; it is important to remember this, whatever variant terminologies and even divergent concepts as to the nature of these entities may exist elsewhere—*e.g.*, as internal powers of the divine organism (Gnostic point of view), as hierarchically ordered intermediaries between the infinite and the finite (Neoplatonic concept), or simply as instruments of the divine activity, neither partaking of the divine substance nor being outside it.

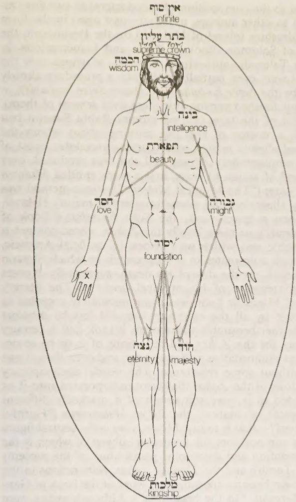
The classical list of the *sefirot* is:

1. *keter 'Elyon*	The Supreme Crown (its identity or nonidentity with the Infinite, *En Sof*, the unknowable deity, remains problematical)
2. *ḥokhma*	Wisdom, the location of primordial ideas in God
3. *bina*	Intelligence, the organizing principle of the universe
4. *ḥesed*	Love, the attribute of goodness
5. *gevura*	Might, the attribute of severity
6. *tif'eret*	Beauty, the mediating principle between the preceding two
7. *netzaḥ*	Eternity
8. *hod*	Majesty
9. *yesod*	Foundation of all the powers active in God
10. *malkhut*	Kingship, identified with the *Shekhina* ("Presence")

The School of Gerona (Catalonia). The double current of the gnosticizing theosophy of the *Sefer ha-bahir* and the contemplative mysticism of the masters of Languedoc became one in the elaborations it was subject to at the hands of the Kabbalists in Catalonia, where the Jewish community of Gerona was, during the first half of the 13th century, a veritable seat of esoterism. These elaborations followed the same overall lines, though they were at the same time highly diversified, depending on the personal inclinations of each writer. To the school of Gerona belong, among others, masters such as Ezra ben Solomon, Azriel of Gerona, Jacob ben Sheshet, Moses ben Naḥman (or Naḥmanides, *c.* 1195–1270, the famous Talmudist, biblical commentator, and mystical philosopher); their influence on the subsequent course of Jewish mysticism is of fundamental importance. None of them has left a complete synthesis of his theosophy; they expressed themselves, with more or less reserve, by means of commentaries, sermons, polemic or apologetic treatises or, at the most, brief summaries for the noninitiated. It is not impossible, however, to discover through these texts their vision of the world and compare it with the views of the Jewish thinkers who attempted to harmonize the biblical-rabbinical tradition with Greco-Arab philosophy, whether of Neoplatonic or Aristotelian inspiration.

The divine "nothing"

At the base of the Kabbalistic view of the world there is an option of faith: it is by a voluntary decision that the unknowable deity—who is "nothing" or "nothingness" (nonfinite) because he is a fullness of being totally inaccessible to any human cogitation—set into motion the process that leads to the visible world. This concept radically separates Kabbala from the determinism from which the philosophy of the period could not, without internal



Sefirot superimposed on the human body (see text).
From C.D. Ginsburg, *Kabbalah: The Essenes*; Samuel Weiser Inc.

contradictions, free the principle of being. In addition it offers a solution consistent with faith to the problem, highly embarrassing for the philosophers, of creation *ex nihilo* (out of nothing): the paradoxical reinterpretation of the concept of the "nothing" eliminates the original matter coeternal with God and solves the opposition between divine transcendence (remoteness from the world) and immanence (presence in the world); issuing from the unfathomable depth of the deity and called to return to it, the world, visible as well as invisible, is radically separated from God, who is at the same time constantly present. The correspondence between the *sefirot*, which are modes of the divine manifestation, and all the degrees of being gives meaning to the structure of the world and to the history of humanity centred on the revelation especially given to the chosen people, a revelation that is a rule of life for this people and, consequently, the criterion of merit and sin, or good and evil. Thus, from the top to the bottom of the ladder, there are but corresponding realities that control one another; contrary to the opinion of the philosophers, evil is also a reality since it is the rupture of the universal harmony. It is also the consequence of this rupture, in the form of punishment, but it is repairable. From this perspective, scrupulous observance of the Torah, the revealed Law (both in the written text and the oral tradition), is the essential factor for the very maintenance of the universe. From that point on, the "rational" motivation of the commandments, which raises insurmountable difficulties for the theologians of philosophical orientation, is in the eyes of the Kabbalists but a false problem; the real problem is the fundamental nature of the Torah. Kabbala brings more than one solution to it, whereas philosophy is not even able to raise the question.

It follows from this general concept that the Jewish faith, with its implications—the conviction of holding the undiluted truth, the faithful preservation of ritual practices, and the eschatological expectation—is safeguarded from all the doubts that either philosophical speculation or the rival religious doctrines of Christianity and Islām could evoke in the minds of Jewish believers. Considered from this point of view, Kabbala, already at the stage it had reached at Gerona, turns out to have been a significant factor in the survival of Judaism, which was exposed everywhere in medieval society to the perils that the history of the period reveals.

Besides the Gerona school and the doctrinal descendants of Isaac the Blind in Languedoc, there was another school of Jewish esoterism in southern Europe during the first half of the 13th century. This school—whose followers preferred to remain anonymous and therefore published their writings, such as the *Sefer ha-ʿiyyun* ("Book of Speculation"), either without giving any author's name or by attributing them to fictitious authorities—directed its speculation both to the highest levels of the divine world, where it discerned further aspects beyond the 10 *sefirot* and attempted to give an idea of them by resorting to the symbolism of light, and to the primordial causes and the archetypes contained in the deity or directly issued from it. The sometimes striking similarity between these speculations and those of John Scotus Erigena, a notable 9th-century Christian philosopher, seems to indicate not only a typological kinship of themes between this Kabbalistic current and Latin-language Christian Neoplatonism but also a concrete influence of the latter upon the former. The same may be true about Isaac the Blind and the school of Gerona, but certain knowledge is lacking.

Sefer ha-temuna. Still another current manifested itself at the same period; it found its literary expression in the *Sefer ha-temuna* ("Book of the Image") of unknown authorship. This very obscure document claims to explain the figures of the letters of the Hebrew alphabet. In fact, the speculation of this treatise bears on two themes that were not foreign to the school of Gerona, but it develops them in a personal manner that decisively influenced the future of Jewish theosophy. On the one hand, it deals with a theory of different cycles through which the world must travel from the time of its emergence to its reabsorption into the primordial unity and, on the other hand, with various readings that correspond to these cycles in the divine manifestation that is constituted by the revealed Scriptures. In other words, the reading, thus the interpretation, and consequently the message of the Torah vary according to the cycles of existence; the passage to a cycle other than that under whose governance humanity is presently living could thus entail the modification, even the abrogation, of the rule of life to which the chosen people are presently subject, an explosive notion that opened the way to an overthrow of the traditional values of Judaism.

The cycles of existence

Medieval German (Ashkenazic) Ḥasidism. The period from *c.* 1150 to 1250, which witnessed the establishment of Kabbala in the south of France and in Spain, is no less important for the shaping of Jewish mysticism in the other branch of European Judaism, in northern France (and England) and in the Rhine and Danube regions of Germany. Unlike medieval Kabbala, which was to experience a broad and varied development starting in the second half of the 13th century, the movement designated somewhat summarily as German (or Ashkenazic, from a biblical place-name conventionally used to designate Germany) Ḥasidism (Pietism), would hardly survive as a living and independent current beyond the second quarter of the 13th century. There was undoubtedly within Franco-German Judaism a certain continuity of mystical tradition, based on the *Sefer Yetzira* and the *Hekhalot* (see above *Sefer Yetzira*); certain elements of theurgy and magic of Babylonian origin had perhaps also reached it through Italy; and it would seem that the gnosticizing current crystallized in the *Sefer ha-bahir* did not pass without leaving traces in Germany. The intellectual atmosphere of Franco-German Judaism, however, differed greatly from that reigning in Spain or even Provence–Languedoc; it was characterized by an almost exclusively Talmudic culture, less intellectual contact with the non-Jewish environment than in the countries of Muslim civilization, and a very limited knowledge of the Jewish theology in Arabic of the Middle East, North Africa, and Spain. This situation would change only in the last third of the 12th century; until then, the "philosophical" equipment of the Franco-German Jewish scholar consisted essentially of a Hebrew paraphrase, dating perhaps from the 11th century, of the treatise *Beliefs and Opinions* by Sa'adia ben Joseph (the great 9th–10th-century Babylonian Jewish scholar and philosopher), and the commentary on the "Book of Creation," written directly in Hebrew (in 946) by the Italian Jew Shabbetai Donnolo. Even when the cultural influence of Spanish Judaism came to be felt more strongly in France–England and Germany, the speculative Kabbala noted above hardly penetrated there. Thinkers within Franco-German Judaism who inclined toward theological speculation had their own problems, which resulted in a mysticism strongly imbued with asceticism, a type of mysticism toward which the general situation of the Jews in those regions contributed, as, especially after the First Crusade, they were severely afflicted by bloody persecutions.

Ashkenazic speculation and asceticism

The main speculative problem was that of the relationship between God in his pure transcendence and total unity and his manifestations in creation, as well as in revelation and communication with inspired men. Reflection on this problem led to the elaboration of various supernatural hierarchies between the inaccessible God and the created universe or the recipient of divine communication; data on angels taken from the Bible and rabbinical and mystical tradition, as well as speculation on the *Shekhina*, were used as material for these hierarchies and also gave a peculiar coloration to liturgical practice. The latter was marked, moreover, by a concern for spiritual concentration by means of fixing the attention on the words and even the letters of the synagogue prayers. Whatever the historical interest of these speculations, they had no great repercussions on the subsequent course of Jewish esoterism; the only exceptions are the mysticism of prayer and demonology, which was sometimes influenced by the beliefs of the Christian environment and fully developed in Ḥasidic circles. On the other hand, the ascetic morality of the movement, which found its literary expression in the work of Eleazar ben Judah of Worms (*c.* 1160–1238) and in the two recensions of the "Book of the Pious" (*Sefer ḥasidim*), was to mark Jewish spirituality, esoteric or not, from then on.

The making of the Zohar (c. 1260–1492). Once the actually marginal episode of German Ḥasidism was finished, almost all of the creative activity in Jewish mysticism was to be situated or would originate in Spain, up to the expulsion of the Jews in 1492.

After the flowering of the schools described above came to an end, around the year 1260, two other currents appeared. The first, in its own manner, resumed relations with Gnosticism in that it placed the problem of evil at the centre of its reflection. The texts that reflect this tendency do not maintain evil in a state of dependence on the "attribute of judgment" within the structure of the *sefirot* set up by the previous Kabbalists but locate it outside the divinity, constructing a parallel system of "left-hand *sefirot*," with a corresponding development of an exuberant demonology. The second movement, whose main representative was the 13th-century visionary-adventurer Abraham ben Samuel Abulafia, found its justification in inner experiences considered "prophetic" and encouraged by training methods akin to those of Yoga, the Byzantine Hesychasts (mystical, quietist monks), and the Muslim Ṣūfīs (mystics); moreover, an important place was given to speculations on the letters and vocalic signs of the Hebrew script. Unlike the protagonists of other mystical schools of Spain that until then had not sought to spread their ideas outside the circle of initiates, Abulafia applied himself in various places to propaganda and exhibitions that disaffected and worried the leaders of Judaism and caused their initiator to be pursued even by the non-Jewish authorities. The numerous writings that he left were later to stimulate a few minds among the Kabbalists.

The world of Moses de León

The work of Moses ben Shem Tov de León, in the last quarter of the 13th century, marked one of the most important turning points in the development of Jewish mysticism. Moses de León was born in the middle of the 13th century and died in 1305; he was the author of several esoteric works, which he signed with his own name. But at the same time, in order to better spread his ideas and to more effectively combat philosophy, which he considered a mortal danger to the Jewish faith, he turned to the composition of pseudepigrapha (writings ascribed to other authors, usually in past ages) in the form of Midrashim (plural of Midrash) on the Pentateuch, the Song of Solomon, Book of Ruth, and Lamentations, in which Talmudic authorities appeared, of whom only the names were even partially authentic, a procedure already used by the *Sefer ha-bahir* (see above *Sefer ha-bahir*); in its most finished version (for there were several of them), the plot of the tales centred around Rabbi Simeon ben Yoḥai, a doctor of the 2nd century, about whom the Talmud already related some curious anecdotes, most of them semilegendary. Moses de León thus produced, over a period of about 30 years, first a work entitled *Midrash ha-ne'elam* ("The Mystical Midrash") whose method was largely allegorical and whose tongue was mainly Hebrew, and then a larger work, the *Sefer ha-zohar* ("Book of Splendour"), or more briefly the *Zohar*, whose content is theosophic and which was written in artificial Aramaic. The book culminates in a long speech in which Simeon ben Yoḥai, on the day of his death, supposedly exposes the quintessence of his mystical doctrine. The literary hoax of Moses de León was not immediately accepted as authentic by all the esoterists and still less by scholars outside the theosophic movement; it took half a century or more for the *Zohar* and imitations of it to be recognized as authoritative ancient works, and even then it was not without some reluctance. The nearly contemporary imitations of the *Zohar* that were incorporated into it or appended to it were sometimes of a markedly different ideological orientation: the *Ra'ya Mehemana* ("Faithful Shepherd"—that is to say, Moses, who is the central figure of this composition, the particular subject of which is the interpretation and theosophic justification of the precepts of the Torah); and the *Tiqqune Zohar*, elaborations in the same vein bearing upon the first word of the book of Genesis (*Bereshit*, "In the beginning"). Although critics were never fully silenced and the authenticity of the *Zohar* was already questioned in the 15th century, the myth created by Moses de León and his imitators became a spiritual reality for the majority of believing Jews; it still retains this character among "Orthodox" Jews. The *Zohar*, believed to be based on supernatural revelations and reinterpreted in diverse ways, would serve as support and reference for all the Jewish theosophies in the centuries ahead.

Doctrine and symbolism of the *Zohar*

As to doctrine, the *Zohar* and its appendixes develop, amplify, and exaggerate speculation and tendencies that already existed, rather than offering any radical innovation. All of the ideas had already been accepted for a long time in Jewish theosophy: the springing forth of being from the depth of the divine "nothing"; the solidarity of the world of the *sefirot* (complicated by the introduction of four ontological levels at each one of which the schema of the 10 *sefirot* is reproduced) with the visible world; the indispensable contribution of man (that is, of the Jew) who observes the biblical and rabbinical precepts in their slightest details, to universal harmony—these emphases remain the main lines of the *Zohar*. But all these themes (the speculations of the *Sefer ha-Temuna*, mentioned above, on the cosmic cycles and the "Prophetic Kabbala" of Abulafia being tacitly set aside) were largely organized and enhanced by the use, or rather the unscrupulous appropriation, of materials taken from rabbinical tradition and ancient esoterism as well as from more recent theological and philosophical currents of thought, despite the lack of esteem that the writers of the Zoharic corpus felt and sought to make others feel toward works created by gentiles.

The method of symbolic representation used by the writings of the Zoharic corpus was supported by a system of interpretation that made use of the originally Christian concept of the fourfold meaning of Scripture: literal, moral, allegorical (philosophical), and mystical. The symbolism thus set up boldly made use of an exuberant anthropomorphic and even erotic imagery whose function was to convey the manifestation of the levels of the *sefirot* to each other and to the extradivine world. The myth of the primordial man (Adam Qadmon), a virtually divine being, reappeared here under a new form, and it was to remain in the subsequent development of Kabbala.

The *Zohar* thus claims to provide a complete explanation of the world, man, history, and the situation of the Jew; on a higher level, to justify the biblical revelation and rabbinical tradition, down to the slightest detail, including the messianic expectation; and thereby to neutralize philosophy. But, while setting itself up as the defender of the traditional religion regulated by the Talmud and its commentaries, in a sense it places itself above tradition, by proclaiming boisterously the incomparable value of the theosophic teaching of "Rabbi Simeon ben Yoḥai" and the superiority of the esoteric doctrine over the Talmudic studies, which were open to all and which, along with the observance of the precepts, were, according to common opinion, supposed to constitute the basic justification of the life of the Jew. There is in this attitude—more accentuated in the *Ra'ya Mehemana* (see above) than in the *Zohar* proper—a revolutionary potentiality, a possible threat to the primacy of practice and study of Torah; the future would show that this danger was not completely unreal.

The Lurianic Kabbala. After the establishment of the Zoharic corpus, no major changes took place in Jewish esoterism until the middle of the 16th century, when in Safed (in Upper Galilee, Palestine; present-day Ẕefat, Israel) a religious centre of extreme importance for Judaism was established, which was mainly inspired by teachers coming from families expelled from Spain. Until the expulsion of the Jews from Spain (1492) and during the two generations that followed it, the Kabbalistic literary output had certainly been abundant, in Spain till the expulsion as well as in Italy and the Middle East; but it was primarily a matter of systematizing or even popularizing the *Zohar* or of extending the speculation already developed in the 13th century; there were also some attempts at reconciling philosophy and Kabbala. It should be noted that even the traditionalist theologians adopted a careful and rather reserved attitude toward Kabbala.

The tragedy for Judaism of the expulsion from Spain and of the forced conversions to Christianity that preceded it by a century, and which would become even more extensive in Portugal shortly afterward, deeply marked the victims. These events, accentuating the already existing pessimism in response to the situation of the Jewish people dispersed among the nations, intensified the messianic expectation. This expectation does not seem to have been unrelated to the beginnings of the printed transmission of Kabbala—the first two printed editions of the *Zohar* date from 1558. All these factors, joined with certain internal developments of speculative Kabbala in the 15th century, prepared the ground for the new theosophy inaugurated by the teaching of Isaac ben Solomon Luria, who was born in Jerusalem in 1534, educated in Egypt, and died in Safed in 1572; although his teaching is traditionally associated with Safed, he spent only the last three years of his life there. Luria wrote very little; his doctrine has been transmitted, amplified, and probably somewhat distorted through the works of his disciples, of which the main one was Ḥayyim Vital (1543–1620), who wrote *'Etz Ḥayyim* ("Tree of Life"), the standard presentation of Lurianic Kabbala.

The teaching of Isaac Luria

The theosophy of Luria, whose novelty was proclaimed by its creator and perfectly realized by the esoterists who held to the Zoharistic Kabbala (organized and codified precisely in Safed, during the lifetime of Luria, by Moses ben Jacob Cordovero, 1522–70), is of extreme complexity in its details, although basically it is but one more attempt to reconcile divine transcendence with immanence and to bring a solution to the problem of evil, which the believer in the divine unity can recognize neither as a power existing independently of God nor as an integral part of him.

The theosophic vision of Luria is expressed in a vast mythical construct, which is typologically akin to certain Gnostic and Manichaean (3rd-century dualistic) systems but which strives at all costs to avoid dualism. The essential elements of this myth are as follows: the withdrawal (*tzimtzum*) executed by the divine light, which originally filled all things, in order to make room for the extradivine; the sinking, as a result of a catastrophic event that occurred during this process, of luminous particles into matter (*qelippot*, "shells," a term already used in Kabbala to designate the evil powers); whence the necessity of saving these particles and returning them to their origin, by means of "repair" or "restoration" (*tiqqun*). This must be the work of the Jew who not only lives in complete conformity to the religious duties imposed on him by tradition but who also dedicates himself, in the framework of a strict asceticism, to a contemplative life founded on mystical prayer and the directed meditation (*kawwana*) of the liturgy, which is supposed to further the harmony (*yiḥud*, "unification") of the innumerable attributes within the divine life. The successive reincarnations of the soul, a constant theme of Kabbala that Lurianism developed and made more complex, are also invested with an important function in the work of "repair." In short, Lurianism proclaims the absolute requirement of an intense mystical life with, as its negative side, an unceasing struggle against the powers of evil. Thus it presents a myth that symbolizes the origin of the world, its fall, and its redemption; it gives meaning to the existence and to the hopes of the Jew, not merely exhorting him to a patient surrender to God but moving him to a redeeming activism, which is the measure of his sanctity. Obviously, such requirements make the ideal of Lurianism possible only for a small elite; ultimately, it is realizable only through the exceptional personage of the "just"—the ideal holy Jew described above.

Shabbetaianism. During the 60 years that followed the death of the founder, the Kabbala linked to the name of Luria and overlaid with accretions from the other mysticisms of Safed spread through the Jewish Diaspora and deeply permeated its spiritual life, liturgy, and devotional practices. It emphasized the necessity of "repair" of a world in which the uneasiness of the Jew kept growing, for in spite of certain favourable factors—the relative tolerance of the Ottoman Empire and the peaceable establishment of an important Marrano (Iberian Jewish, or Sefardic) community in Amsterdam—there was no overall solution to the problem of the "conversos" (converts) who had remained in the Iberian Peninsula. The other half of the Jewish people, the Ashkenazim, also experienced a serious crisis: its most prosperous and dynamic section, the Jewish population of Poland, was sorely tried, almost totally ruined, and in large part forced to move back toward the west because of the massacres and the destruction that took place during the Cossack uprising of 1648. These ideological and historical data may provide the necessary context for understanding the astonishing though short-lived success of Rabbi Shabbetai Tzevi of Smyrna (1626–76), who proclaimed himself messiah in 1665. The "Messiah" was forcibly converted to Islām in 1666 and ended his life in exile 10 years later, but despite his failure he had faithful followers. A sect was thus born and survived largely thanks to the activity of Nathan of Gaza (*c.* 1644–90), an unwearying propagandist for the "Messiah," who justified the actions of Shabbetai Tzevi, which were contrary to the Law, and his final apostasy by theories that were based on the Lurian theory of "repair": it had to be understood as the descent of the just into the abyss of the "shells" in order to liberate from it the captive particles of divine light. The Shabbetaian crisis lasted nearly a century, some of its aftereffects even longer. It led to the formation of sects whose members were externally converted to Islām—*e.g.*, the Dönme (Turkish, "apostates") of Salonika, whose descendants still live in Turkey—or to Roman Catholicism—*e.g.*, the Polish supporters of Jacob Frank (1726–91), the self-proclaimed Messiah and Catholic convert. In Bohemia–Moravia, however, the Frankists outwardly remained Jews. This crisis did not discredit Kabbala, but it led the spiritual authorities of Judaism to watch over and severely curtail its spread and to exercise rigorous ideological control, by concrete acts of censorship and repression, over anyone, even a person of tested piety and recognized knowledge, who was suspected of Shabbetaian sympathies or of messianic pretensions.

Shabbetai Tzevi, a 17th-century "Messiah"

Modern Ḥasidism. Though it is true that the messianic movement centred around Shabbetai Tzevi could only produce disillusionment and that if it had not been contained it could have led Judaism to its ruin, yet it answered not only the theosophic aspirations of a small number of visionary scholars but also an affective need of the Jewish masses that was left unsatisfied by the dry intellectualism of the Talmudists and the economic and social oppression of the ruling classes (both Jewish and non-Jewish). This was the case especially in Poland, which before the partition of the Polish kingdom (1772–95) included Lithuanian, Belorussian, and Ukrainian territories. It was there that the so-called Ḥasidic movement, in no way connected with medieval German Ḥasidism, originated around the middle of the 18th century—a movement in which the Lurian Kabbala, theoretically maintained as the basis of speculation, underwent adjustments and transformations that continue to the present day.

If modern Ḥasidism may be regarded as a mass movement, having a minimum of organization, using the methods of propaganda and preaching, and forming groups of acknowledged members, then the legend is credible that traces it back to a single founder, Israel ben Eliezer, known as Baʿal Shem Ṭov (Master of the Good Name; that is, a possessor—he was not the only one of his kind—of the secret of the ineffable name of God, which bestows an infallible power to heal and perform other miraculous operations). This man was born about 1700 and died in 1760 in southern Poland. Though relatively untrained according to the norms of the rabbinical Judaism of his time, he was a spiritual personage of exceptional quality and was able to win to his ideas not only the common people but also many representatives of the intellectual elite. The mist of legend that surrounds him is too dense for it to be possible to reconstruct entirely his personal doctrine, which he probably never systematized. Drawing his inspiration from the methods of the itinerant preachers whose activity was becoming more intense in 18th-century eastern European Judaism, he delivered his teaching in the form of homiletic interpretations of sacred texts, having recourse to fables and parables borrowed from daily life and from folklore; this method remained constant in Ḥasidism, but it is undeniably an exaggeration and even an error in perspective to consider, as did Martin Buber (see below *Modern Jewish mysticism*), that the tale and the anecdote are the most authentic expression of the doctrine and the spirituality of Ḥasidism. It is indeed in the doctrinal works, most of them expressed in the form of sermons on the weekly sections of the Pentateuch and other liturgical lessons, that the thought of the Ḥasidic "rabbis" is expressed. It is very diversified thought, for there are as many bodies of doctrine in Ḥasidism as there were creative spirits during the first three generations of the movement. It is, nevertheless, possible to point to a few traits that are fundamental and common to Ḥasidism as a whole.

The Baʿal Shem Ṭov

In theory, it remains rooted in the Lurianic Kabbala—and nothing essential separates it at this point from its most implacable adversaries in the traditional Judaism of eastern Europe. What is unique to it is to have made of *devequt,* "being-with-God," an object of aspiration and even a constant duty for all Jews and in all circumstances of life, even those seemingly most profane; in other words, it demands a total spiritualization of Jewish existence. This requirement entails a reevaluation, less new in its principle than in its concrete application, of the speculative concepts of Kabbala: the emphasis is placed on the inner life of the believer, and it is at this level that the supercosmic drama is played (a drama whose stage was, according to bookish theosophy, in the universe of the *sefirot*); according to several teachers, the same emphasis on inwardness holds for messianic redemption. At the same time, Ḥasidism transforms into social reality a requirement that was also part of the Lurian doctrine of "repair," though it was unfortunately distorted by Shabbetaianism: Ḥasidism puts at the centre of the religious life and organization of the group, as an indispensible guide and unquestioned authority, the inspired leader, endowed with supernatural powers—the "just" (*tzaddiq*), the "miracle-working rabbi" (*Wunder-rebbe*). Ḥasidism thus produced, wherever it triumphed, an undeniable spiritual renewal; the reverse of the medal was the cult of personality, competition between "dynasties" of "rabbis," obstinacy in maintaining the Ḥasidic community apart from the surrounding society, with all the social and economic consequences that this will to isolation entailed and for which it would be false to lay all the blame on the environment, despite its definite hostility toward the Jews.

The *tzaddiq*

From its very beginnings, Ḥasidism was to encounter strong resistance on the part of the official Judaism of the period, which had been sensitized to the anarchism of the Shabbetaians and was at the same time solicitous for the prerogatives of the established community leaders and rabbis, the vigilant guardians over the traditional laws and their application, who were confined to the formal study of the Talmud and its commentaries. The behaviour of the followers of Ḥasidism, though irreproachable in its strict, even rigorous observance of ritual rules, displayed several traits that were distasteful to its adversaries (besides the unconditional submission to the *tzaddiq,* who often doubled as the rabbi of the official congregation): desertion of the general communal synagogues, meetings in small conventicles, modifications of the liturgy, casual dress during prayer, and preference given to mystical meditation rather than to the dialectical study of the Talmud, which requires instead serious intellectual concentration. Nevertheless, the conflict between the Ḥasidim and the "Opponents" (Mitnaggedim) did not finally degenerate into a schism; after three generations, a kind of tacit compromise was established between the two tendencies—Ḥasidic and Talmudic—without the consciousness of differences ever being erased. The compromise was rather to the advantage of Ḥasidism, but not without a few concessions on its part, notably on the question of education.

The strong organization of the Ḥasidic groups allowed them to survive the dislocation of eastern European Judaism as a result of the events of World War II, but its vital centres are today in the United States rather than in Palestine, in part because of economic reasons, in part because of the more or less reserved, and at sometimes frankly hostile, attitude of the Ḥasidic "rabbis" toward political Zionism and the State of Israel. The best known of the U.S.-based groups is the very active Lubavitchers (after Lyubavichi, Russia, seat of a famous school of Ḥasidism), whose headquarters are in the Crown Heights district of Brooklyn, New York.

MODERN JEWISH MYSTICISM

The role played by Kabbala and Ḥasidism in the thought and spirituality of contemporary Judaism is far from being insignificant, though its importance is not as great as in former times. Of course, there is hardly any really living Kabbalistic and Ḥasidic literature, but the personal thought of religious writers such as Abraham Isaac Kook (*c.* 1865–1935), spiritual leader, mystic, and chief rabbi of Palestine, continues to exercise a marked influence. Furthermore, the renewal of religious thought in "westernized" Jewish circles between the two wars received a powerful impulse from the philosopher Martin Buber (1878–1965), whose work is in part devoted to the propagation of Ḥasidic ideology as he understood it. "Neo-Orthodoxy," founded in Germany by Samson Raphael Hirsch (1808–88), was quite indifferent to mysticism at the outset, but it too came to be influenced by it, especially after the rediscovery of living Judaism in Poland during World War I by Western Jewish thinkers. Also significant is the work of Abraham Joshua Heschel (1907–72), a Polish Jewish writer of distinguished Ḥasidic background and double culture—traditional and Western—who emigrated to the United States.

Christian and secular Kabbala

Jewish mysticism also has exercised some influence on thought outside the Jewish community. Kabbala, distorted and deflected from its own intentions, transcended the frontiers of Judaism and helped nourish and stimulate certain currents of thought in Christian society, from the Renaissance to the present: "Christian Kabbala," born in the 15th century under the impetus of Jewish converts from Spain and Italy, claimed to find in the Kabbalistic documents, touched up if necessary or even forged, arguments for the truths of the Christian faith. Thus a certain number of Christian Humanist scholars became interested in Jewish mysticism and several of them acquired a fairly extensive knowledge of it on the basis of authentic texts. Among them were Giovanni Pico della Mirandola (1463–

94) and Gilles of Viterbo (Egidio da Viterbo; *c.* 1465–1532) in Italy, and Johannes Reuchlin (1455–1522), who was responsible for writing one of the principal expositions of Kabbala in a language accessible to the learned non-Jewish public (*De arte Cabbalistica*, 1517), in Germany, while the visionary Guillaume Postel (1510–81) was attracting disciples in France. The occult philosophy of the 16th century, the "natural philosophy" of the 17th and 18th centuries, and the occult and theosophic theories that are cultivated even today and that have coloured the ideology of Freemasonry—all of these focus and continue to make borrowings from Kabbala, though they rarely grasp its spirit and meaning. The same is true of most of the books on Kabbala put out by publishers of occult and theosophic literature today.

The rigorous scholarly study of Jewish mysticism is a very recent phenomenon. The state of mind and the tendencies of the founders of the "science of Judaism" (the scholarly study of Jewish religion, literature, history, etc.) in Germany during the first half of the 19th century were too permeated with rationalism to be favourable to scholarly investigation of a movement judged to be obscurantist and retrograde. Granting some valuable earlier works, research on a large scale and application of the proved methods of philology and history of religions began only with the work of Gershom G. Scholem, who was professor of Kabbala at the Hebrew University, Jerusalem from 1923 to 1965, and has been continued by his disciples, both direct and indirect. This research touched all of the areas of Jewish mysticism that are briefly described in this article; however, the gaps in knowledge remain serious in every area. Critical editions of mystical texts are few in number; unpublished documents are cataloged in a very incomplete manner; and only a few monographs on writers and particular themes exist, though these are indispensable preliminaries to a detailed and thorough synthesis. It is to be hoped that the one outlined by Scholem in 1941, in his *Major Trends in Jewish Mysticism*, though of exceptional value in its time, will be taken up again and completed. (G.V.)

Jewish myth and legend

Jewish myth and legend comprises a vast body of stories transmitted over the past 3,000 years in Hebrew and in vernacular dialects, such as Yiddish (Judeo-German) and Ladino (Judeo-Spanish), spoken by Jews in various parts of the world. These stories have played an important role in the history of Jewish religion and culture.

SIGNIFICANCE AND CHARACTERISTICS

Apart from their intrinsic appeal, Jewish myths and legends claim attention for three special reasons: (1) Those incorporated in the Old Testament now form part and parcel of the cultural heritage of the Western world and have exerted a profound influence on its literature and art. (2) During the Middle Ages Jews were among the principal transmitters of Oriental tales to the West, so that many familiar Eastern stories can be traced to Jewish compilations. (3) Since these stories have been accumulated through centuries of constant migration, they provide an unrivalled body of "clinical" material for studying the process by which popular tales in fact travel and are transformed.

Basic character and role

Not all of the stories are of Jewish origin; many can be readily paralleled elsewhere and are derived from tales the Jews picked up from their non-Jewish neighbours in the lands of their dispersion. Even what is borrowed, however, is usually impressed with a distinctive Jewish stamp, being adapted to point up some precept of the Jewish religion, to illustrate some facet of Jewish life, or to exemplify some trait of Jewish character and temperament. The dominant overall feature of the stories is, indeed, their religious and moral tone; most of them are, in fact, told specifically as part of the homiletic exposition of Scripture. Such stories are taught to Jews from early childhood as a regular part of their religious education. To the tradition-minded Jew, therefore, they are more than mere literary fancies and assume a kind of doctrinal complexion. Biblical characters and events present themselves to him more in the lineaments of later legend than in their original biblical form; while popular notions about heaven and hell, rewards and punishments, the coming of the Messiah, and the resurrection of the dead derive mainly from this source rather than from Scripture itself.

Virtually all the standard types of folktale are represented. Conspicuously absent, however, are pure fairy tales because fairies, elves, and the like are foreign to the Jewish imagination, which prefers to people the otherworld with angels and demons subservient to God.

A distinction must be made, of course, between myth and legend. In common parlance, a myth is a story about gods or otherworldly beings. Judaism, however, is a rigorously monotheistic religion; hence, in this narrower sense, there can be no original Jewish myths. Nevertheless, from the earliest times, Jews have not disdained to borrow those of their pagan neighbours and then adapt them to their own religious outlook. If, however, the term is interpreted in a larger sense, to mean the portrayal of continuous, transtemporal concerns in the context of particular and punctual events, myth is indeed one of the essential vehicles by which Judaism conveys its message; for it is only when historical happenings are translated into this wider dimension that they cease to be mere antiquarian data and acquire continuing relevance. In Judaism, for example, the Exodus from Egypt is projected mythically from something that happened at a particular time into something that is continually happening, and it thus comes to exemplify the situation and experience of all men everywhere—their emergence from the bondage of obscurantism, their individual revelations at their individual Sinais, their trek through a figurative wilderness, even their death in it so that their children or children's children may eventually reach the figurative "promised land." By the same token, the historical destruction of the Temple of Jerusalem is transformed by myth into a paradigm of the continuing mutual estrangement of God and man, their exile from one another.

Legend, on the other hand, implies no more than a fanciful embroidering of purportedly historical fact. Unlike myth, it does not transcend the punctual and local.

SOURCES AND DEVELOPMENT

Myth and legend in the Old Testament. The vast repertoire of Jewish myths and legends begins with the Old Testament. Their overall purpose in Scripture is to illustrate the ways of God with man, as exemplified both in historical events and in personal experience. The stories themselves are often derived from current popular lore and possess abundant parallels in other cultures, both ancient and modern. In each case, however, they are given a peculiar and distinctive twist.

Myths in Genesis

Myths. Old Testament myths are found mainly in the first 11 chapters of Genesis, the first book of the Bible. They are concerned with the creation of the world and of man, the origin of the continuing human condition, the primeval Deluge, the distribution of peoples, and the variation of languages.

The basic stories are derived from the popular lore of the ancient Middle East and can be paralleled in the extant literature of the peoples of the area. The Mesopotamians, for instance, also knew of an earthly paradise such as Eden, and the figure of the cherubim—properly griffins rather than nightgowned angels—was known to the Canaanites. In the Bible, however, this mythical garden of the gods becomes the scene of man's fall and the background of a story designed to account for the natural limitations of human life. Similarly, the Babylonians, too, told of the formation of man from clay, but in the scriptural version his function is to bear rule over all other creatures, whereas in the pagan tale it is to serve as an earthly menial of the gods. Again, the story of the Deluge, including the elements of the ark and the dispatch of the raven and dove, appears already in the Babylonian myths of Gilgamesh and Atrahasis. There, however, the hero is eventually made immortal, whereas in the Bible this detail is omitted because to the Israelite mind no child of woman could receive that status. Lastly, while the story of the Tower of Babel was told originally to account for the stepped temples (ziggurats) of Babylonia, to the Hebrew writer its purpose is simply to inculcate the moral lesson that man should not build beyond his assigned station.

Myths in the Prophets and Holy Writings

Scattered through the Prophets and Holy Writings (the two latter portions of the Hebrew Bible) are allusions to other ancient myths—*e.g.*, to that of a primordial combat between Yahweh and a monster variously named Leviathan (Wriggly), Rahab (Braggart), or simply Sir Sea or Dragon. The Babylonians told likewise of a fight between their god Marduk and the monster Tiamat; the Hittites told of a battle between the weather god and the dragon Illuyankas; while from Ras Shamra (ancient Ugarit), in north Syria, has come a Canaanite poem relating the discomfiture of Sir Sea by the deity Baal and the rout of an opponent named Leviathan. (Originally, this myth probably referred to the annual subjugation of the floods.)

Ancient myths are utilized also in the form of passing allusions or poetic "conceits," much as modern Westerners may speak of Cupid or the Muses. Thus, there are references in the prophetic books to a celestial upstart hurled to Earth on account of his brashness and to the imprisonment of certain rebellious constellations.

The prophets used such myths paradigmatically to illustrate the hand of God in contemporary events or to reinforce their forecasts. Thus, to Isaiah the primeval dragon becomes the symbol of that continuous force of chaos and evil that will again have to be vanquished before the Kingdom of God can be established on Earth. Similarly, for Ezekiel the celestial upstart serves as the prototype of the prince of Tyre, destined for an imminent fall; and Habakkuk sees in the impending rout of certain invaders a repetition on the stage of history of Yahweh's mythical sortie against the monster of the sea.

Legends and other tales. Old Testament legends often embellish the accounts of national heroes with standard motifs drawn from popular lore. Thus, the story (in Genesis) of Joseph and Potiphar's wife recurs substantially (with other characters) in an Egyptian papyrus of the 13th century BCE. The depositing of the infant Moses in the bulrushes (in Exodus) has an earlier counterpart in a Babylonian tale about Sargon, king of Akkad (*c.* 2334–*c.* 2279 BCE), and is paralleled later in legends associated with the Persian Cyrus and with Tu-Küeh, the fabled founder of the Turkish nation. Jephthah's rash vow (in Judges) whereby he is committed to sacrifice his daughter recalls the classical legend of Idomeneus of Crete, who had similarly to slay his own son. The motif of the letter whereby David engineers the death in battle of Bathsheba's husband recurs in Homer's story of Bellerophon and again in the episode of Rosencrantz and Guildenstern in *Hamlet.* The celebrated judgment of Solomon concerning the child claimed by two contending women is told, albeit with variations of detail, about Buddha, Confucius, and other Oriental sages; while the story of how Jonah was swallowed by a "great fish" but subsequently disgorged intact finds a parallel in the Indian tale of the hero Śaktideva, who experienced the same thing during his quest for the Golden City. On the other hand, it should be observed that many of the parallels commonly cited from the folklore of primitive peoples may be, in fact, mere playbacks of biblical material picked up from Christian missionaries.

Etiologic tales

Sometimes, worldwide folktales serve in the Old Testament to account for the names of places in Palestine or for the origins of traditional customs and institutions. Thus, the familiar story of the man who has to struggle with the personified current of a river before he can cross it is localized (in Genesis) at the ford of Jabbok simply because that name suggests the Hebrew word *abḳ* ("struggle"); and Samson's felling of 1,000 Philistines with the jawbone of an ass is placed at Ramath-leḥi because *leḥi* is Hebrew for "jawbone." Similarly, a taboo against eating the sciatic nerve of an animal is validated (in Genesis) by the legend that Jacob was struck in the hip when he tussled with an otherworldly being at Penuel (Face of God); and the custom of annually bewailing the vanished spirit of fertility is rationalized (in Judges) as a lamentation for the hapless daughter of Jephthah.

Besides myths and legends the Old Testament also contains a few examples of fables (didactic tales in which animals or plants play human roles). Thus, the serpent in Eden talks to Eve, and Balaam's ass not only speaks but also "flairs" spirits; while in the celebrated parable of Jotham (in Judges) trees compete for kingship.

Finally, in the Book of Job (38:31) there are allusions to star myths concerning the binding of Orion (called the Fool) and the "chaining" of the Pleiades.

Contemporary interpretations. The tendency to interpret biblical tales and legends as authentic historical records or as allegories, or as the relics of solar, lunar, and astral myths, is now a thing of the past. For the modern folklorist, their primary interest lies in the fact that they push back to remote antiquity several tales and motifs long known from later literature. For the theologian, however, they pose the deeper problem of distinguishing clearly between the permanent message of Scripture and the particular form in which it is conveyed. Such a process of "demythologization" is today one of the central concerns of religious thought. It involves recognition of the fact that the natural language of religious truth is myth so that the continuing relevance of ancient scriptures depends not on a total rejection of that vehicle but rather on a constant expansion and remodelling of it—*i.e.*, on remythologization rather than demythologization. In the final analysis, the traditional portrayal of God himself is simply a mythical representation of ultimate reality, but that reality transcends the particular images in which it happens to be expressed. At the same time, it must be clearly understood that expressions that can be reconciled with modern Western patterns of thought only if taken as metaphors were literal statements of fact to ancient and primitive peoples. Gods, for example, were not merely "personifications" of natural phenomena but rather the effective potencies of the phenomena themselves conceived from the start as personal beings, much as a modern child might conceive of a railroad engine as "Mr. Choochoo."

Myth and legend in the Persian period. When, in 539 BCE, the Jews came under Persian domination, they absorbed a good deal of Iranian folklore about spirits and demons, the eventual dissolution of the world in a fiery ordeal, and its eventual renewal. This introduced a new element into Jewish myth and legend. Hierarchies of angels, archangels such as Michael, Gabriel, and Uriel (modelled loosely upon the six Iranian spiritual entities, the *amesha spentas*), and the demonic figures of Satan, Belial, and Asmodeus (corresponding to the Iranian Angra Mainyu [Ahriman], Druj, and Aēshma daeva) now entered their popular mythology, and there was a preoccupation with apocalyptic visions of heaven and hell and of the Last Days. Unfortunately, no Jewish texts of this genre from the Persian period itself are extant so that these new elements can be recognized only inferentially from their survival in later times, notably in such products of the ensuing Hellenistic age as the Dead Sea Scrolls.

Angelology, demonology, and eschatology

The principal monument of Jewish story in the Persian period is the biblical Book of Esther, and this is basically the Judaized version of a Persian novella about the shrewdness of harem queens. The story was adapted to account for a popular festival named Purim, but this is probably a transmogrification of the Persian New Year. Such leading elements of the tale as the parade of Mordecai through the streets dressed in royal robes, the fight between the Jews and their adversaries, and the hanging of Haman and his sons seem, indeed, to reflect customs associated with that occasion, viz., the ceremonial ride of a common citizen through the capital, the mock combat between two teams representing Old Year and New Year, and the execution of the Old Year in effigy.

Myth and legend in the Hellenistic period. *Historiated Bibles and legendary histories.* When, in 330 BCE, Alexander the Great completed his conquest of the Middle East, Judaism entered a new phase. The dominant features of the ensuing Hellenistic age were an increasing cosmopolitanism and a fusion of Oriental and Greek cultures. These found expression in Jewish myth and legend in the composition (in Greek) of stories designed to link the Bible with general history, to correlate biblical and Greek leg-

"Esther and Ahasuerus," tempera painting of Konrad Witz, 15th century. In the Öffentliche Kunstsammlung Basel. 85.5 × 79.5 cm.

By courtesy of the Offentliche Kunstsammlung Basel, Switzerland

ends, and to claim for the Hebrew patriarchs a major role in the development of the arts and sciences. It was asserted, for instance, that Abraham had taught astrology to the king of Egypt; that his and Keturah's sons had aided Heracles against the giant Antaeus; and that Moses, blithely identified both with the semi-mythical Greek poet Musaeus and with the Egyptian Thoth, had been the teacher of Orpheus (putative founder of one of the then current "mystery cults") and the inventor of navigation, architecture, and the hieroglyphic script. Leading writers in this vein were Artapanus, Eupolemus, and Cleodemus (all *c.* 100 BCE), but their works are known to us only from stray quotations by Eusebius and Clement of Alexandria, early Church Fathers.

Novelistic versions of biblical figures

Furthermore, the Jews followed a current Greek literary fashion of retelling Homeric and other ancient legends in "modernized," novelistic versions, well seasoned with romantic elaborations. Among the Dead Sea Scrolls has been found a paraphrase of Genesis in which the biblical narrative is tricked out with several familiar folklore motifs. Thus, when Noah is born, the house is filled with light, just as it is said elsewhere to have been at the birth of the Roman king Servius Tullius, of Buddha, and (later) of several Christian saints. When Abraham's life is threatened he dreams of a cedar about to be felled—the same omen said to have presaged the deaths of Domitian and Severus Alexander. (True, the parallels are of later date, but they illustrate the persistence of age-old popular traditions.) The same trend toward fanciful elaboration of scriptural tales is manifested also in the *Testaments of the Twelve Patriarchs* ("testaments" meaning last wills), in which the virtues and weaknesses of the sons of Jacob are illustrated by moralistic legends. There is also a lengthy paraphrase of early biblical narratives, mistakenly attributed to Philo, the famous Alexandrian Jewish philosopher of the first century CE.

Apocrypha and Pseudepigrapha. The principal monuments of Jewish literature during the Hellenistic period are the works known collectively as the Apocrypha and Pseudepigrapha. The former are certain later writings excluded by Jews from the canon of the Old Testament but found in the Greek Septuagint version. The latter are other late writings not included in any authorized version of the Scriptures and spuriously attributed to biblical personalities.

Judaized versions of foreign tales

The Apocrypha include several Judaized versions of tales well represented in other cultures. The book of Tobit, for instance, turns largely on the widespread motifs of "The Grateful Dead" and the "Demon in the Bridal Chamber." The former relates how a traveller who gives burial to a dishonoured corpse is subsequently aided by a chance companion who turns out to be the spirit of the deceased. The latter tells how a succession of bridegrooms die on the nuptial night through the presence of a demon beside the bridal bed. Similarly, in Bel and the Dragon (2nd century BCE) occurs the equally familiar motif that fraud (in this case perpetrated in a temple) is detected by the imprint of the culprit's foot on strewn ashes—a motif that reappears later in the French and Celtic romance of Tristan and Iseult. Again, *Susanna and the Elders* (also 2nd century BCE) revolves around the well-worn theme that a charge of unchastity levelled against a beautiful woman is refuted when a clever youngster ("Daniel come to judgment") points out discrepancies in the testimony of her accusers. The story has a close parallel in a Samaritan tale about the daughter of a high priest in the 1st century CE; while the motif of the clever youngster who surpasses seasoned judges recurs later in infancy gospels and in the tale of ʿAlī Khamājah in *The Thousand and One Nights.*

The most interesting folktale in the Pseudepigrapha is that contained in *The Martyrdom of Isaiah* (1st century CE?), which tells how the prophet, fleeing from King Manasseh, hid in a tree that opened miraculously and how he eventually perished when it was sawn asunder. A similar tale is related in the Talmud about a certain Isaac ben Joseph and (later) in the Persian epic *Shāh-nāmeh* (*c.* 1000 CE) about the hero Jamshīd.

Myth and legend in Talmud and Midrash. *Midrash and Haggada.* Toward the end of the 1st century CE, through a process known as "canonization," certain traditional Hebrew writings came to be recognized as an authoritative corpus of divine revelation, later called the Hebrew Bible or Old Testament. The study of them became, henceforth, an essential element of the Jewish religion. This meant that the sacred text had to be subjected to a form of interpretation that would bring out its universal significance and permanent relevance. The process was known as Midrash (literally "searching the Scriptures"), and a leading constituent of it was the spicing of homiletic discourses with elaborative legends—a pedagogic device called Haggada ("storytelling"). Originally transmitted orally, the legends were eventually committed to writing in that vast sea of literature known as the Talmud (the authoritative compendium of early rabbinic law and lore), as well as in later compilations geared to particular books or sections of the Old Testament, to scriptural lessons read in the services of the synagogue, or to specific biblical characters or moral themes (see also above *Torah*).

The range of Haggada is virtually inexhaustible; a few representative examples must suffice. In regard to biblical characters, both Moses and David were born circumcised; Cain had a twin sister; Abraham will sit at the gate of hell to reproach the damned on Judgment Day; Aaron once locked the angel of death in the tabernacle; Solomon understood the language of animals; King Hiram, who supplied materials for the Temple, entered paradise alive; the flesh of Leviathan will feed the righteous in the world to come.

Indebtedness of Haggada to non-Jewish sources

In such fanciful elaborations of Scriptures, Haggada does not disdain to draw on classical tales. The men of Sodom, it is said, subjected itinerant strangers to the ordeal of Procrustes' bed; the Earth opened to rescue newborn Hebrew males from the Pharaoh, as it did for Amphiaraus, the prophet of Argos, when he fled from Periclymenus after the attack on Thebes; Moses spoke at birth, as did Apollo; Solomon's ring, cast into the river, was retrieved from a fish that had swallowed it, as was that of Polycrates, the tyrant of Samos, in the story told by Herodotus; the Queen of Sheba had the feet of an ass, like the child-stealing witch (Onoskelis) of Greek folklore; no rain ever fell on the altar at Jerusalem, just as none was said to have fallen on Mt. Olympus.

Other familiar motifs also appear. Moses qualifies as a husband for Zipporah by alone being able to pluck a rod from Jethro's garden—a variant of the tale told later about the sword Excalibur in the Arthurian legend; David's harp is played at night by the wind, like that of Aeolus; Isaiah, like Achilles and Siegfried, has only one vulnerable spot in his body—his mouth; Job has a magic belt, which relieves his pains.

Legends are developed also from fanciful interpretations of scriptural verses. Thus, Adam is said to have fallen only a few hours after his creation because the Hebrew text of Ps. 49:12 can be literally rendered "Adam does not last the night in glory." Lamech slays the wandering Cain—a fanciful interpretation of his boast in Gen. 4:23–24. Melchizedek is immortal in view of Ps. 110:4: "You are a priest for ever after the order of Melchizedek." The first man is a hermaphrodite (this notion has analogues elsewhere) because Gen. 1:27 says of God's creation, "Male and female he created them."

Fables and animal stories. Midrash also uses fables paralleled in non-Jewish sources. Aesop's fable of the "Lion and the Crane" is quoted by a rabbi of the 1st century CE, and the tales of the "Fox in the Vineyard" and of the "Camel Who Got Slit Ears for Wanting Horns" likewise make their appearance.

Sometimes, too, material is drawn from medieval bestiaries (manuals on animals, real or imaginary, with symbolic or moralistic interpretations). Bears, we are told, lack mother's milk; hares and hyenas can change sex; only one pair of unicorns exists at a time; there is a gigantic bird (*ziz*) that reaches from Earth to sky.

Contribution of Haggada to Christian and Islāmic legends. Several of the stories related in Haggadic literature were later adopted and adapted by Christian writers. Thus, the legend that Adam was created out of virgin soil was taken to prefigure the fact that the second Adam (*i.e.,* Jesus) was likewise born of a virgin; while the story that the soil in question was taken from the site of the future Temple was transformed into the claim that Adam had been molded out of the dust of Calvary. Similarly, the legend that, at the dedication of the Temple, the doors had swung open automatically to admit the ark of the Covenant was transferred to the consecration of a church by St. Basil; and the Talmudic tale that the bronze Nicanor gates of the Temple had floated to Jerusalem when cast overboard for ballast during their shipment from Alexandria was applied to the doors of a sacred edifice erected in honour of St. Giles.

Nor was it only the Christians who absorbed Haggadic legends. The Qur'ān, the sacred book of Islām, likewise incorporates a good deal of such material in its treatment of such biblical characters as Joseph, Moses, David, and Solomon.

Myth and legend in the medieval period. *Jewish contribution to diffusion of folktales.* The Middle Ages was a singularly productive period in the history of Jewish myth and legend. Jews now began to play a prominent role in the transmission of Oriental tales to the West and thereby enhanced their own repertoire with a goodly amount of secular material. Especially in Spain and Italy, Arabic versions of standard collections were translated into Hebrew and thence into Latin, thus spreading the stories to the Christian world. The Indic fables of Bidpai, for example, were rendered into Hebrew from the 8th-century Arabic version of ʿAbd Allāh ibn al-Muqaffaʿ, and from this Hebrew rendering there subsequently developed, in the 12th century, John of Capua's *Directorium humanae vitae* ("Guide for Human Life"), one of the most celebrated repertoires of moralistic tales (*exempla*) used by Christian preachers. So, too, the famous *Senbād-nāmeh* ("Fables of Sinbad"; one of the sources, incidentally, of Boccaccio's *Decameron*) was rendered from Arabic into Hebrew and thence into Latin; while the renowned romance of *Barlaam and Josaphat*—itself a Christian adaptation of tales about the Buddha—found its Jewish counterpart in a compilation entitled *The Prince and the Dervish,* adapted, from an Arabic text, by Abraham ben Samuel ibn Ḥisdai, a leader of Spanish Jewry in the 13th century.

Hebrew versions of medieval romances. Here, too, however, the traffic moved in both directions: Hebrew translations were also made from Latin and other European languages. There are, for instance, several Hebrew adaptations of the *Alexander Romance,* based mainly (though not exclusively) on Leo of Naples' Latin rendering of the Greek original by Callisthenes. The central theme is, of course, the exploits of the great Macedonian conqueror, and the narrative is spiced with fanciful accounts of his adventures in foreign lands and of the outlandish peoples he encounters. There is likewise a Hebrew reworking of the Arthurian legend, in the form of a secular sermon in which Arthurian and biblical scenes are blithely mixed together. Finally, there is a Hebrew *Ysopet* (the common title for a medieval version of Aesop) that shares several of its fables with the famous collection made by Marie de France in the late 12th century.

The *Alexander Romance,* Arthurian legend, and Aesop's fables

Jewish contributions to Christian and Islāmic tales. Moreover, apart from these Hebrew translations of Oriental and European works, a good deal of earlier haggadic material is embodied in the *Disciplina clericalis* of Peter Alfonsi, a baptized Jew of Aragon originally known as Moses Sephardi. This book, composed in the 12th century, is the oldest European collection of novellas and served as a primary source for the celebrated *Gesta Romanorum* ("Deeds of the Romans") of the same period—a major quarry for European storytellers, poets, and dramatists for many centuries.

Haggadic material percolated also to Arabic writers during this period. Not only does the Qur'ān incorporate such material but also the Egyptian recension of *The Thousand and One Nights* seems to have drawn extensively on Jewish sources, as, for instance, in its tales of "The Sultan and His Three Sons," "The Angel of Death," "Alexander and the Pious Man," and the legend of Baliqiyah.

Major medieval Hebrew collections. Between the 11th and 13th century the tendency developed in Europe to compile, both for entertainment and edification, comprehensive collections of tales and fables; standard examples are the British *Gesta Romanorum,* the Spanish *El novellino,* and the aforementioned *Disciplina clericalis.* Among Jews similar collections were made, especially in Morocco as well as in Moorish Spain. Two of the most important are *The Book of Comfort* by Nissim ben Jacob ben Nissim of al-Qayrawān (11th century) and *The Book of Delight* by Joseph ben Meir ibn Zabara of Spain (end of the 12th century). The former, composed in Judeo-Arabic, is a collection of some 60 moralizing tales designed to comfort the author's father-in-law on the loss of a son. It belongs to a well-known genre of Arabic literature, derives mainly from Arabic sources, and is permeated by a preoccupation with divine justice, typical of the Muʿtazilite school of Islāmic theology. It was later translated into Hebrew. *The Book of Delight* consists of 15 tales, largely about the wiles of women, exchanged between two travelling companions—a form of cadre, or "enclosing tale," adopted on a more extensive scale by Chaucer in his *Canterbury Tales,* which dates from the same period. Typical is the tale of the "Silversmith and His Wife," which relates how a craftsman, persuaded by his greedy wife to make a statue of a princess, gets his hands cut off by the king for violating the Islāmic law against making images, while his wife reaps rich rewards from the flattered princess. Although most of the stories are taken from Arabic sources, some indeed find parallels in rabbinic literature To the latter category belongs, for instance, the famous tale of the matron of Ephesus, who, while keeping vigil over her husband's tomb, at the same time engages in an intrigue with a guard posted nearby to watch over the corpses of certain crucified robbers. When, during one of their trysts, one of the corpses is stolen and her lover therefore faces punishment, the shrewd woman exhumes the body of her husband and substitutes it. This tale is found already in the *Satyricon* of Petronius and was later used by Voltaire in his *Zadig* and by the 20th-century English playwright Christopher Fry in his *A Phoenix Too Frequent.*

The story of the matron of Ephesus

Of the same genre but deriving mainly from west European rather than Arabic sources are the *Mishle shuʿalim* ("Fox Fables") of Berechiah ha-Nakdan (the Punctuator), who may have lived in England toward the end of the 12th century. About half of these tales recur in Marie de France's *Ysopet,* and only one of them is of specifically Jewish origin. Berechiah's work was translated into Latin and thence became a favourite repertoire of European storytellers.

Among anonymous compendiums of this type is *The*

Alphabet of Ben Sira, extant in two recensions, probably of the 11th century. This is basically a collection of proverbs attributed to the famous sage of the apocryphal book Ecclesiasticus (Wisdom of Jesus the Son of Sirach). In one of the recensions they are illustrated by appropriate tales. The author is represented as an infant prodigy who performs much the same feats of sapience as are attributed to Jesus in some of the Infancy Gospels.

Medieval historiated Bibles and legendary histories. Two other developments mark the history of Jewish myth and legend during the Middle Ages. The first was a revival of the Hellenistic vogue of compiling large-scale compendiums in which the history of the Jews was "integrated," in legendary fashion, with that of the world in general and especially with classical traditions. Two major works of this kind, both composed (apparently) in Italy during the 9th century, are (1) *Josippon*, composed by a certain Ben Gorion, which presents a fanciful record from the creation onward and contains numerous references to foreign nations; and (2) the *Book of Jashar*, a colourful account from Adam to Joshua, named for the ancient book of heroic songs and sagas mentioned in the Bible (Josh. 10:13; II Sam. 1:18). There is also a voluminous *Chronicles of Jerahmeel*, written in the Rhineland in the 14th century. This draws largely on Pseudo-Philo's earlier compilation, mentioned above, and is of special interest because it includes Hebrew and Aramaic versions of certain books of the Apocrypha.

The *Josippon*, *Jashar*, and *Jerahmeel* collections

Medieval Haggadic compendiums. The other development was the gathering of Haggadic legends and tales into comprehensive, systematic compendiums. Works of this kind are (1) *Yalquṭ Shimeʿoni* ("The Collection of Simeon"), attributed to a certain Rabbi Simeon of Frankfurt am Main; (2) *Midrash ha-gadol* ("The Great Midrash"), composed after the death of Moses Maimonides (1204), whom it quotes; and (3) the *Midrash of David ha-Nagid*, grandson of Maimonides. About 100 years later appeared a similar work, *Yalquṭ ha-Makiri* ("The Collection of Makhir"), on the Prophets and Holy Writings, compiled by one Makhir ben Abba Mari in Spain (see above *Torah*). It has been suggested that the compilation of such works was spurred by the necessity of providing "ammunition" for the public disputations with Christian ecclesiastics that the church forced upon Jewish scholars in this period.

Myth and legend in the modern period. *Kabbalistic tales.* In the 16th century, Jewish myth and legend took several new directions. The disappointment of messianic expectations through the dismal eclipse of the pretender Shabbetai Tzevi produced, by way of compensation, an increased interest in occult speculation and in the mystical lore of the Kabbala (esoteric Jewish mysticism). Important schools of Kabbala arose in Italy and at Safed, in Palestine, and tales of the miraculous Faust-like powers of such masters as Isaac Luria and Ḥayyim Vital Calabrese began to circulate freely after their deaths.

The legend of Joseph della Reyna

Another reaction to the dashing of messianic hopes is represented by the beautiful story of the Kabbalist Joseph della Reyna and his five disciples, who go journeying through the world to oust Satan and prepare the way for the Deliverer. Warned by the spirits of such worthies as Rabbi Simeon ben Yoḥai and the prophet Elijah, they nevertheless succeed eventually in procuring their blessing and help and are sent on to the angel Metatron. The latter furnishes them with protective spells and spices and advises Joseph to inscribe the ineffable name of God on a metal plate. When, however, they reach the end of their journey Satan and his wife, Lilith, attack them in the form of huge dogs. When the dogs are subdued they beg for food. Moved to pity, Joseph gives them spices to revive them. At once they summon a host of devils. Two of the disciples die of terror; two go mad, and only Joseph and one disciple are left. The Messiah weeps in heaven, and Elijah hides the great horn of salvation. A voice rings out telling Joseph that it is vain to attempt to hasten the footsteps of the Redeemer.

The repertory of Jewish tales and legends was seasoned, however, by other elements. During the 16th century—the age of the great navigators—stories began to circulate about the discovery of the Ten Lost Tribes in remote parts of the world.

Judeo-German (Yiddish) tales. It was at the same period that Judeo-German (Yiddish) came increasingly to replace Hebrew as the language of Jewish tales and legends in Europe, a major factor in this development being the desire to render them accessible to women unschooled in the sacred tongue. Not only were the synagogal lessons from Scripture legendarily embellished in a so-called *Taitsh Humesh* ("Yiddish Pentateuch"), in the more fancifully titled *Tzeʾena u-reʾena* ("Go Forth and See"; *cf.* S. of Sol. 3:11) by Jacob ben Isaac Ashkenazi, and in adaptations of the story of Esther designed for dramatic presentation on the feast of Purim, but the Hebrew *Chronicles of Josippon* also assumed Yiddish dress. More secular productions were a verse rendition of the Arthurian legend, entitled *Artus Hof* ("The Court of King Arthur"), based largely on Gravenberg's medieval *Wigalois*, and the *Bove Buch* by Elijah Levita, which retold the romance of Sir Bevis of Southampton.

These "frivolous" productions were in time offset by collections of moral and ethical tales. The principal of these are (1) the *Brantspiegel*, attributed to a certain Moses Henoch (Prague 1572), and (2) the *Maʿaseh Buch* ("Story Book"), a compendium of 254 tales compiled by Jacob ben Abraham of Meseritz and first published at Basel in 1602. The latter was drawn mainly from the Talmud but was supplemented by later legends about medieval rabbis. Jewish legends also circulated in the form of ephemeral chapbooks, a large selection of which is preserved in the library of the Yiddish Scientific Institute in New York City.

Judeo-Persian and Judeo-Spanish (Ladino) tales. A similar development, though on a lesser scale, took place among Jews who spoke other vernacular dialects. Major monuments of Judeo-Persian literature are poetic embellishments of biblical narratives composed by a certain Shāhīn of Shīrāz in the 14th century and by Joseph ben Isaac Yahudi (*i.e.*, the Jew) some 300 years later. These, however, are exercises in virtuosity rather than in creative storytelling. In Judeo-Spanish (Ladino) there are versified elaborations of the story of Joseph, entitled *Coplas de Yoçef* ("Song of Joseph"), composed, in 1732, by Abraham de Toledo and embodying a certain amount of traditional haggadic material. From a revival of literary activity in the 18th century comes a comprehensive "legendary Bible" called Me-ʿam Loʿez ("From a People of Strange Tongue"; *cf.* Ps. 114:1), begun by one Jacob Culi and continued by later writers, as well as several renderings of standard Hebrew collections and a number of Purim plays. Until the Nazi holocaust in the 1940s, Judeo-Spanish folktales were still current in Macedonia and Yugoslavia, but these leaned more on Balkan than on Jewish sources.

Ḥasidic tales. The rise of the Ḥasidic sect (a popular pietistic-mystical movement) in eastern Europe at the end of the 18th century begat a host of legends (circulated mainly through chapbooks) concerning the lives, wise sayings, and miracles of such *tzaddiqim*, or masters, as Israel ben Eliezer, "the Besht" (1700–60), and Dov Baer of Meseritz (died 1772). (See also above *Jewish mysticism*.) These, however, are anecdotes rather than formally structured stories and often borrow from non-Jewish sources.

Droll stories. To the popular creativity of the ghetto belong also the droll tales of the Wise Men of Chełm (in Poland)—Jewish counterparts of the German noodles (stupid people; hence "noodle stories") of Schildburg and of the more familiar English Wise Men of Gotham. These, too, were circulated mainly in Yiddish popular prints. Typical of them is the tale of the two "sages" who went for a walk, one carrying an umbrella and the other without one. Suddenly it began to rain. "Open your umbrella," said the one without one. "It won't help," answered the other, "it's full of holes." "Then why did you bring it?" rejoined his friend. "I didn't think it would rain," was the reply.

Modern Israeli folktales. The gathering of Jews from many lands into the modern state of Israel has made that country a happy hunting ground for the student of Jewish folktales. Assiduous work has been undertaken by Dov Noy of Hebrew University in Jerusalem, aided by enthusiastic amateurs throughout the country. Mainly, however, the stories are retellings of traditional material. (T.H.G.)

Judaism in world perspective

RELATION WITH NON-JUDAIC RELIGIONS

Exclusivist and universalist emphases. The biblical tradition out of which Judaism emerged was predominantly exclusivist ("no other gods"). The gods of the nations were regarded as "no gods" and their worshippers as deluded, while the God of Israel was acclaimed as the sole lord of history, and the Creator of heaven and earth. The unexpected universalist implications of this exclusivism are most forcibly expressed in an oft-quoted verse from Amos (9:7):

God's rule over all the nations

> "Are you not like the Ethiopians to me, O people of Israel?" says the Lord. "Did I not bring up Israel from the land of Egypt, and the Philistines from Caphtor and the Syrians from Kir?"

Here the universal rule of the God of Israel is unmistakably proclaimed. Yet in the same book (3:1–2), after referring to the deliverance from Egypt—an act recognized as similar to that occurring in the affairs of other peoples—the prophet, speaking for God, says: "You only have I known of all the families of the earth." Thus the exclusivism has two focuses, one universal, the other particularistic. The ultimate claim of the universalistic position is found in Malachi 1:11: "For from the rising of the sun to its setting my name is great among the nations." This, however, in no way negates the special covenantal relationship between God and his people; indeed, it is this universalistic theme that underscores that special relation. To interpret Judaism's stance toward other religious systems in any other way is to fail to do justice to its inner dialectic. It is neither a bland latitudinarianism that admits any or all viewpoints and practices, nor a fanatical intolerance but rather a subtle interplay of affirmation and rejection. The latter is directed primarily against the worship of finite things or aspects—idolatry—the basic failure of the peoples who are the objects of the same divine solicitude as is Israel. If the religions of the nations are rejected because of their failure fully and truly to know God, the peoples themselves are not. Living under the covenant with Noah (see above), their fulfillment of such responsibilities provides for their acceptance, for they are not expected to live within the realm of Torah (see also *Relations with other religions* below).

Relation to Christianity. Judaism's relation to Christianity is a complicated one because of the close historical interconnections between them. From a Judaic standpoint, Christianity is or was a Jewish "heresy" and as such may be judged somewhat differently than other religions. Its claims over against Judaism as the true fulfillment of the covenant and, thus, as the true Israel have given rise throughout the centuries to polemics of varying intensity. The rise to power of the church and the embodiment of its anti-Judaic sentiments and attitudes in the political structures and processes of Christian nations made sharply negative Jewish responses inevitable. Nevertheless, during the Middle Ages Jewish thinkers attempted to avoid designating Christianity as idolatry and even to argue that, in a special way, being derived from Judaism, it was fulfilling—at least on the moral plane—the divine purpose.

Christian and Jewish polemic and counter-polemic

In modern times the relation has undergone changes necessitated by the newer situations into which the Jewish community has moved. This does not mean that the polemical-apologetic stance has come entirely to an end. The rejection of Judaism as a living religion by Christians continued and continues, argued not so much on dogmatic as on scholarly grounds. The Jewish response to this has often been countercriticism. Beyond this, however, there has been a growing inclination within the Jewish community to respond to the development of an affirmative theology of Judaism in both the Roman Catholic and Protestant churches by providing a theology of Christianity within Jewish thought. Occasional formulations in this direction have appeared, but it is far too early to know exactly what will emerge. At the same time, it must be noted, there are those who see no need for such a movement, arguing that the failure of the Christian churches in recent years to respond adequately to the tragedies of Jewish existence precludes any real engagement of one with the other.

Relation to Islām. The emergence of Islām in Arabia in the 7th century CE brought Judaism face to face with a second religious movement that derived some of its ideas and structures from the older tradition. In this case, as in that of Christianity, the new religion claimed a special relation with Judaism. Muḥammad held that the faith he proclaimed was none other than the pristine religion of Abraham, the father of Ishmael—progenitor of the Arabs—as well as of Isaac, from whom the people of Israel descended. That religion had been distorted both by Judaism and Christianity; and Muḥammad, the "seal" of the prophets, had been called by God to restore it to its purity. The confrontation between Judaism and Islām, as that with Christianity, was coloured by political and social considerations both before and after Islām moved out of Arabia to build a world empire (including the conquest and settlement of Palestine). During the subsequent period, the intellectual development of the Islāmic world and the emergence of theologians and philosophers of the highest order challenged Judaism and had considerable influence on the rise of similar thinkers within that community. Given the strong monotheism and the anti-iconic attitude of Islām, many of the questions that arose between Judaism and trinitarian and iconic Christianity were not an issue between Judaism and Islām. The crucial point of dispute here was the nature of prophecy, given Muḥammad's claim concerning his culminating role in the prophetic tradition. The medieval period thus saw polemics directed against that claim and, as in the case of the theological work of Moses Maimonides, *More nevukhim* (*The Guide of the Perplexed*), an exposition of the nature of prophecy that, without directly dealing with Muḥammad's claim, may be understood to undercut it. Nonetheless, Islām, too, was understood to contribute to the fulfillment of the divine purpose. From the late medieval period onward, the intellectual engagement between the two religions diminished with the general decline in the Turkish Empire that then embraced the Muslim world. In modern times it has not yet been renewed for many reasons. Once the political problems in the eastern Mediterranean between the State of Israel and the Arab world have been meliorated, the contiguity of the two communities suggests an inevitable renewal of conversations on the religious as on many other levels.

Relations with other religions. Judaism's encounters with religions other than Christianity and Islām have been in large measure limited to the past. In the Hellenistic world, it confronted and rejected the varieties of syncretistic cults that grew up. Within the Sāsānian Empire it was forced to deal with Zoroastrianism, but the outlines of its response have not yet been entirely disentangled from the literature of the period. In the modern world, particularly in the most recent period, it has come face to face with the religions of the Middle and Far East, but beyond a few tentative explorations nothing tangible has appeared. What seems certain is that, considering the growing interest in and exchange between East and West, Jewish thinkers will not be able to rest with older formulations concerning the nature of other religious systems. Without compromising its own faith or falling into an uncritical relativism, Judaism may indeed in the future seek a new way of understanding and relating to the varieties of religious systems facing it on the world scene.

Judaism confronts the religions of mankind

THE ROLE OF JUDAISM IN WESTERN CULTURE AND CIVILIZATION

Its historic role. Given the relationship between Judaism and Christianity—the dominant religious force in the development of Western culture—the role of Judaism in that development was significant. Although the church drew from other sources as well, its retention of the sacred Scriptures of the synagogue (the "Old Testament") as an integral part of its Bible—a decision sharply debated in the 2nd century CE—was crucial. Not only was the development of its ideas and doctrines deeply influenced, but it received as well an ethical dynamism that constantly overcame an inclination to withdraw into world-denying isolation. It was, however, not only Judaism's heritage but its persistence that touched Western civilization. The continuing existence of the Jews, even as a pariah people, was both a challenge and a warning; and ultimately, at the beginning of the modern era, their liberation from the shackles of discrimination, segregation, and rejection was understood by many to be the touchstone of all human liberty. Until the final ghettoization of the Jew—it is well to remember that the term "ghetto" belongs in the first instance to Jewish history—at the end of the Middle Ages and the beginning of the Renaissance, intellectual contact between Judaism and Christianity, and thus with Western culture, did not cease. Jerome translated the Hebrew Bible into Latin with the aid of Jewish scholars; Luther, into German with the aid of commentaries beholden to Jewish authors. Jewish thinkers mediated the remarkable intellectual achievements of the Islāmic world to Christian Europe and added their own contributions as well. Even heresies within the church found, on occasion, their inspiration or prototype in Judaism.

Its present role. In the modern world, while the influence of Jews has increased in almost every realm of cultural life, the impact of Judaism has diminished. The reason for this is not difficult to find. The Gentile leaders who extended emancipation to the Jews at the end of the 18th and beginning of the 19th centuries, while eager to grant political equality to the individual Jew, did so with the implicit and explicit requirement that conformity through reforms of Judaism be agreed to. With the transformation of Judaism into an ecclesiastical institution, largely on the model of German Protestant churches, its ideas and structures took on the cast of its environment in a way quite unlike what had ensued in its earlier confrontations with various philosophical systems. Indeed, for some, Judaism and 19th-century European thought were held to be not merely congruent but identical. Thus, while numerous contributors to diverse aspects of Western culture and civilization are to be found among Jews of the 20th century—scientists, politicians, statesmen, scholars, musicians, artists—their activities cannot, except in specific instances, be considered as deriving from Judaism as it has been sketched above.

Future prospects. Two events of the 20th century have, however, confronted Judaism in such ways as to suggest that its wrestling with them and their profound challenge to it may presage a new role and a new influence for Judaism: "Auschwitz" and the establishment of the State of Israel. The premeditated murder of some 6,000,000 European Jews by the Nazis for no other reason than that they were Jews, has shaken Jewish thinkers to their very core. Indeed, so traumatic was this event, that for almost two decades following it, no substantial attempt was made to plumb its meaning. At the same time, the reappearance of the State of Israel, viewed for the most part from outside the Jewish community as nothing more than a political event, has set in motion an entirely different chain of theological inquiry. These two happenings have clearly, but in as yet unpredictable ways, begun to work and to move within the thought of contemporary Judaism. Out of this working and moving there may emerge an inescapable spiritual impact upon Western culture and civilization, which have, as yet, resolutely refused to face the realities these fateful occurrences represent. If contemporary Judaism is able to say what they mean, however haltingly, it will have renewed its potent relationship to the Western world, and, given the nature of contemporary society, established a similar bond with the Eastern world as well. (L.H.S.)

The two key events: Auschwitz and the establishment of Israel

BIBLIOGRAPHY

General history: SALO W. BARON, *A Social and Religious History of the Jews,* 2nd ed., 15 vol. (1952–73), a comprehensive presentation of the intertwined social and religious history with copious bibliographical information critically evaluated; LOUIS FINKELSTEIN (ed.), *The Jews: Their History, Culture and Religion,* 4th ed., 3 vol. (1970–71), critical essays by outstanding authorities on the major aspects of Jewish history and culture; JULIUS GUTTMANN, *Die Philosophie des Judentums* (1933; Eng. trans., *Philosophies of Judaism,* 1964), the best single volume on the history of Jewish thought from ancient times to the present, especially valuable for medieval Jewish philosophy; LEO W. SCHWARZ (ed.), *Great Ages and Ideas of the Jewish People* (1956), interpretive and highly readable essays by six historians on Jewish history, with emphasis on intellectual history, intended primarily for the layman; GERSHOM G. SCHOLEM, *Major Trends in Jewish Mysticism,* 3rd rev. ed. (1954), the classic treatment of mystical doctrines and schools in Judaism; MAX L. MARGOLIS and ALEXANDER MARX, *A History of the Jewish People* (1927, reprinted 1958), an excellent, readable, introductory survey. See also ROBERT M. SELTZER, *Jewish People, Jewish Thought: The Jewish Experience in History* (1980).

Biblical Judaism: (*General reference*): *The Interpreter's Dictionary of the Bible,* 4 vol. (1962). (*Surveys of the culture and religion of ancient Israel*): JOHANNES PEDERSEN, *Israel: Its Life and Culture,* 4 vol. in 2 (1926–40, reprinted 1959); W.F. ALBRIGHT, *From the Stone Age to Christianity,* 2nd ed. (1957); YEHEZKEL KAUFMANN, *The Religion of Israel, from Its Beginnings to the Babylonian Exile* (1960) and *The Babylonian Captivity and Deutero-Isaiah* (1970); ROLAND DE VAUX, *Les Institutions de l'Ancien Testament,* 2 vol. (1958–60; Eng. trans., *Ancient Israel: Its Life and Institutions,* 1961); GERHARD VON RAD, *Theologie des Alten Testaments,* 2nd ed. (1958; Eng. trans., *Old Testament Theology,* 2 vol., 1962–65); HELMER RINGGREN, *Israelitische Religion* (1963; Eng. trans., 1966). (*Annual bibliographic keys*): *Elenchus Bibliographicus Biblicus,* ed. by P. NOBER (1920–); *Book List* of the British Society for Old Testament Study. (*Special topics*): E.A. SPEISER, "The Biblical Idea of History in Its Common Near Eastern Setting," *Oriental and Biblical Studies: Collected Writings of E.A. Speiser,* ed. by J.J. FINKELSTEIN and MOSHE GREENBERG, pp. 187–210 (1967); MENAHEM HARAN, "The Religion of the Patriarchs: Beliefs and Practices," *Patriarchs: The World History of the Jewish People,* vol. 2, pp. 219–245 (1970); MOSHE GREENBERG, "Crimes and Punishments," *Interpreter's Dictionary of the Bible,* vol. 1, pp. 733–744 (1962); "Some Postulates of Biblical Criminal Law," *The Jewish Expression,* ed. by JUDAH GOLDIN, pp. 18–37 (1970); AELRED CODY, *A History of Old Testament Priesthood* (1969); JOHANNES LINDBLOM, *Prophecy in Ancient Israel* (1962); E.J. BICKERMAN, "The Historical Foundations of Postbiblical Judaism," *The Jews: Their History, Culture, and Religion,* ed. by LOUIS FINKELSTEIN, 4th ed., vol. 1, pp. 70–114 (1960); DAN JACOBSON, *The Story of the Stories: The Chosen People and Its God* (1982).

Hellenistic Judaism: (*Bibliographies*): GERHARD DELLING (ed.), *Bibliographie zur jüdisch-hellenistischen und intertestamentarischen Literatur 1900–1970,* 2nd ed. (1975), extremely comprehensive bibliography on religion and literature of Diaspora Judaism, arranged according to topics, with separate bibliographies on every major Hellenistic Jewish author and on each book of the Apocrypha and Pseudepigrapha; RALPH MARCUS, "A Selected Bibliography (1920–1945) of the Jews in the Hellenistic-Roman Period," *Proceedings of the American Academy for Jewish Research,* 16:97–181 (1946–47), covers both Palestine and the Diaspora, helpful for noting works that are useful introductions and works that are indispensable to the specialist; LOUIS H. FELDMAN, *Scholarship on Philo and Josephus, 1937–1962* (1963), critical bibliography, arranged topically, with comments on both Diaspora and Palestinian Judaism generally. (*Papyrological and archaeological sourcebooks*): VICTOR A. TCHERIKOVER, ALEXANDER FUKS, and MENAHEM STERN (eds.), *Corpus Papyrorum Judaicarum,* 3 vol. (1957–64), contains text, translation, bibliography, and commentary on all papyri and inscriptions pertaining to Jews from 323 BCE to 641 CE—thoroughly reliable; ERWIN R. GOODENOUGH, *Jewish Symbols in the Greco-Roman Period,* 13 vol. (1953–68), a magnificent, exhaustive collection of the archaeological findings, with highly insightful, if controversial, commentary. (*Standard scholarly treatments of Hellenistic Judaism*): VICTOR A. TCHERIKOVER, *Hellenistic Civilization and the Jews* (1959; orig. pub. in Hebrew, 1930), extremely meticulous and generally balanced (though with an anti-theological bias), particularly in dealing with the political and social factors in both Palestinian and Diaspora Jewry; ROBERT H. PFEIFFER, *History of New Testament Times, with an Introduction to the Apocrypha* (1949), a sane and useful, if unoriginal, survey of the political, religious, and literary history of Palestinian and especially Diaspora Judaism, from 200 BCE to 200 CE; MOSES HADAS, *Hellenistic Culture: Fusion and Diffusion* (1959), highly suggestive, though often extravagant, treatment of the interaction of Hellenism and other cultures, especially Judaism. (*Introductory popular treatments of Hellenistic Judaism*): RALPH MARCUS, "The Hellenistic Age," in LEO W. SCHWARZ (ed.), *Great Ages and Ideas of the Jewish People,* pp. 93–139 (1956), extremely judicious and readable account by an eminent authority; for a readable guide to the literature, together with representative samples, see his "Hellenistic Jewish Literature," in LOUIS FINKELSTEIN (ed.), *The Jews: Their History, Culture and Religion,* 3rd. ed., vol. 2, pp. 1077–1115 (1960). (*Works on Palestinian Judaism*): GEORGE FOOT

MOORE, *Judaism in the First Centuries of the Christian Era*, 3 vol. (1927–30, reprinted 1966–67), classic treatment based primarily on the Talmudic corpus, though the view of a Pharisaic "normative" Judaism has since been strongly challenged; SOLOMON ZEITLIN, *The Rise and Fall of the Judaean State: A Political, Social and Religious History of the Second Commonwealth*, 2 vol. (1962–67), stimulating and often highly original survey of the period from 332 BCE to 66 CE, though the scholarly dogmatism is occasionally jarring; SAUL LIEBERMAN, *Greek in Jewish Palestine*, 2nd ed. (1965) and *Hellenism in Jewish Palestine*, 2nd ed. (1962), highly significant, ingenious, and learned illustrations of the influence of Greek culture on the language and exegetical format of the Palestinian rabbis. (*Works on Diaspora Judaism*): HARRY A. WOLFSON, *Philo: Foundations of Religious Philosophy in Judaism, Christianity, and Islam*, rev. ed., 2 vol. (1962), a great and seminal, though controversial, work that makes Alexandrian Judaism a collateral branch of Palestinian Pharisaic Judaism; LOUIS H. FELDMAN, "The Orthodoxy of the Jews in Hellenistic Egypt," *Jewish Social Studies*, 22:215–237 (1960), a survey using literature, papyri, and art objects to examine the synthesis of Greek culture and Judaism in the upper and lower classes, respectively, of Hellenistic Egypt. See also JOHN J. COLLINS, *Between Athens and Jerusalem: Jewish Identity in the Hellenistic Diaspora* (1983); and MENAHEM MOR and URIEL RAPPAPORT, *Bibliography of Works on Jewish History in the Hellenistic and Roman Periods* (1982).

Rabbinic Judaism: (*Palestinian Judaism*): SOLOMON SCHECHTER, *Aspects of Rabbinic Theology* (1961), a concise, authoritative, and engaging treatment of classical (*i.e.*, rabbinic) Judaism. (*Babylonian Judaism*): JACOB NEUSNER, *A History of the Jews in Babylonia*, 5 vol. (1965–70; vol. 1, rev. 1969), the most comprehensive treatment of Babylonian Jewry during the Tannaitic and Amoraic periods. (*Judeo-Arabic culture*): S.D. GOITEIN, *Jews and Arabs: Their Contacts Through the Ages*, 3rd rev. ed. (1974), a popular work by the ranking authority on all aspects of Jewish–Arabic symbiosis, particularly valuable for the medieval period; ABRAHAM IBN DAUD, *Sefer ha-Qabbalah* (*The Book of Tradition*), ed. and trans. by GERSON D. COHEN (1967), the classic medieval Hebrew chronicle with analytic essays on Spanish Jewry's "golden age." (*Jews of medieval Europe*): ISRAEL ABRAHAMS, *Jewish Life in the Middle Ages*, new ed. rev. by CECIL ROTH (1932), delightful and erudite studies of social life and institutions in medieval Europe; MORITZ GÜDEMANN, *Geschichte des Erziehungswesens und der Cultur der Juden*, 3 vol. (1880–88), a social and intellectual history of medieval Ashkenazic Jewry (France, Germany, Italy); CECIL ROTH, *The Jews in the Renaissance* (1959), lucid and informative, but with little critical analysis, valuable on Jewish contact with Christian men of letters; GERSHOM SCHOLEM, *Shabbethai Zevi*, 2 vol. (1957), a penetrating and comprehensive study (in Hebrew) of the great false Messiah as well as of his religious antecedents and legacy; STEPHEN SHAROT, *Messianism, Mysticism, and Magic: A Sociological Analysis of Jewish Religious Movements* (1982), an informative scholarly study.

Contemporary Judaism: The most convenient summary for the study of modern Jewish history is HOWARD MORLEY SACHAR, *The Course of Modern Jewish History*, updated and expanded ed. (1977). Modern Jewish thought and movements are covered in a useful manual by JOSEPH BLAU, *Modern Varieties of Judaism* (1966). NATHAN ROTENSTREICH, *Jewish Philosophy in Modern Times* (1968); and the last (modern) section of JULIUS GUTTMAN, *Philosophies of Judaism* (1964), are more advanced. The very best book on Jewish religion in the modern age is unfortunately still untranslated: MAX WIENER, *Jüdische Religion im Zeitalter der Emanzipation* (1933). For Zionism, see the only attempt at a comprehensive reader in English, ARTHUR HERTZBERG (ed.), *The Zionist Idea* (1959). There are two excellent expositions of Judaism from a Reform–Liberal point of view: LEON ROTH, *Judaism: A Portrait* (1960); and LEO BAECK, *The Essence of Judaism* (1961). Conservative Judaism is well described in a book about its early history, MOSHE DAVIS, *The Emergence of Conservative Judaism* (1963); and by JACOB AGUS, *Dialogue and Tradition* (1969), the essays of a distinguished Conservative thinker. Reconstructionism is best understood in the words of its founder, MORDECAI M. KAPLAN, *Judaism As a Civilization*, 2nd ed. (1957). The standard modern single volume about Orthodox Judaism is ISIDORE EPSTEIN, *Judaism* (1935). Neo-Ḥasidism was best described by its greatest exponent, MARTIN BUBER, *The Origin and Meaning of Hasidism* (1960). Later studies include EUGENE B. BOROWITZ, *Choices in Modern Jewish Thought: A Partisan Guide* (1983); PAUL R. MENDES-FLOHR and JEHUDA REINHARZ (eds.), *The Jew in the Modern World* (1980); DAN ROSS, *Acts of Faith: A Journey to the Fringes of Jewish Identity* (1982).

Judaic literature: B. GERHARDSSON, *Memory and Manuscript* (1961), contains a description of the methods and techniques by which the oral tradition was transmitted. H.L. STRACK, *Introduction to the Talmud and Midrash* (1931); and M. MIELZINER, *Introduction to the Talmud*, 4th ed. (1968), are still the best introductions for the general reader. The latter is particularly helpful in explaining Talmudic dialectic terminology and debate. J. BOWKER, *The Targums and Rabbinic Literature* (1969); and E. DEUTSCH, *The Talmud* (1895), are both descriptions of the Talmud, the former concentrating upon Talmudic literary compilations and the latter upon Talmudic content. The introduction of J. GOLDIN, *The Living Talmud* (1957), contains a vivid description of Talmudic debate. C. ALBECK, *Introduction to the Talmud*, in Hebrew (1969); and Z.H. CHAJES, *The Student's Guide Through the Talmud* (Eng. trans. 1952), are more advanced introductions, the former analytical and scientific and the latter representing the traditional view. J. NEUSNER (ed.), *The Formation of the Babylonian Talmud* (1970), contains summaries of some research by modern scholars on the question of how the Talmud was formed. L. GINZBERG, "Introduction to the Talmud," in *A Commentary on the Palestinian Talmud*, vol. 1 (1941), is the only introduction to the Palestinian Talmud available in the English language. L. FINKELSTEIN, *Akiba* (1962), is a historical and sociological approach to the development of Halakha. See SAUL LIEBERMAN, *op. cit.*, for the interrelationship between the ancient rabbinic world and its Gentile environment. B. COHEN, *Jewish and Roman Law*, 2 vol. (1966); and I. HERZOG, *The Main Institutions of Jewish Law*, 2 vol. (1966–67), are the best English descriptions of Jewish law. J. Z. LAUTERBACH, *Rabbinic Essays* (1951); and E.E. URBACH, *The Sages: Their Concepts and Beliefs*, in Hebrew (1969), cover the major aspects of rabbinic theology. I. HEINEMANN, *Paths of the Aggadah*, in Hebrew (1970); M. KADUSHIN, *The Rabbinic Mind*, 3rd ed. (1972); and D. BEN AMOS, *Narrative Forms of the Haggadah: Structural Analysis* (1969), discuss Haggadic methods, forms, concepts, and thinking. L. ZUNZ, *Die Gottesdienstlichen Vorträge der Juden* (1892; updated Hebrew translation, 1950), is a thorough historical survey of Haggadic literature. D. NOY, *Motif-Index of the Talmudic-Midrashic literature* (1954); J.J. SLOTKI, *Index Volume to the Soncino Talmud* (1952); M. GASTER, *The Exempla of the Rabbis* (1924, rev. ed. 1968); and C.G. MONTEFIORE and H. LOEWE, *A Rabbinic Anthology* (1938), are very helpful as reference guides; while W.O.E. OESTERLEY, H. LOEWE, and E.I.J. ROSENTHAL, *Judaism and Christianity*, rev. ed. (1969); and C. MERCHAVIA, *The Church Versus Talmudic and Midrashic Literature*, in Hebrew (1970), describe the relationship between the church and rabbinic Judaism. E.R. BEVAN and C. SINGER (eds.), *The Legacy of Israel* (1927), deals with the influence of Judaism on world culture. Extensive bibliographies may be found in the works of Gerhardsson, Mielziner, Bowker, and Ben Amos. JACOB NEUSNER, *Judaism: The Evidence of the Mishnah* (1981), introduces new methods of textual criticism.

Basic beliefs and doctrines: KAUFMANN KOHLER, *Jewish Theology: Systematically and Historically Considered* (1918, reprinted 1928); YEHEZKEL KAUFMANN, *The Religion of Israel from Its Beginnings to the Babylonian Exile* (1960; orig. pub. in Hebrew, 1937–56), an abridgement and translation of the work of one of the most influential Jewish biblical scholars of modern times; GEORGE FOOT MOORE, *Judaism in the First Centuries of the Christian Era*, 3 vol. (1927–30), a masterful work by a distinguished Christian student of Judaism; SOLOMON SCHECHTER, *Some Aspects of Rabbinic Theology* (1909, reprinted 1936 and 1961), an insightful presentation of the basic doctrines; CLAUDE G. MONTEFIORE and HERBERT LOEWE, *A Rabbinic Anthology* (1960), a collection of materials from rabbinic sources, arranged under theologic headings, with ample notes and discussions; JULIUS GUTTMANN, *Philosophies of Judaism: The History of Jewish Philosophy from Biblical Times to Franz Rosenzweig* (1964), a philosophy of Judaism in the form of a history of philosophy in Judaism; ARTHUR HERTZBERG (ed.), *Judaism* (1961); JACOB NEUSNER, *The Way of Torah: An Introduction to Judaism*, 3rd ed. (1979), a very helpful statement using a history-of-religions approach; LEO BAECK, *Dieses Volk; jüdische Existenz*, 2 vol. (1955–57; Eng. trans., *This People Israel: The Meaning of Jewish Existence*, 1964), a masterful interpretation of Jewish affirmations set within an historical context; ABRAHAM E. MILLGRAM (ed.), *Great Jewish Ideas* (1964), a collection of essays by various scholars setting forth classic positions, covering a wide range of theological ideas; JACOB B. AGUS, *The Jewish Quest* (1983), a collection of essays on basic concepts of Jewish theology.

Ethics and society: SIMON BERNFELD (comp.), *The Foundations of Jewish Ethics*, 2nd rev. ed. (1968), source materials covering Jewish history, with brief introductions by several notable scholars; MORITZ LAZARUS, *Die Ethik des Judenthums*, 2 vol. (1898–1911; Eng. trans., *The Ethics of Judaism*, 1900), a classic presentation from a 19th-century perspective; SAMUEL S. COHON, *Judaism: A Way of Life* (1948), written from the Reform point of view but deeply sympathetic to a wide range of ideas. Jewish ethics is also explored in ANNE ROIPHE, *Generation Without Memory: A Jewish Journey in Christian Amer-

ica (1981); and illustrated with literary examples in FRANCINE KLAGSBRUN (comp.), *Voices of Wisdom: Jewish Ideals and Ethics for Everyday Living* (1980).

Basic practices and institutions: LEWIS N. DEMBITZ, *Jewish Services in Synagogue and Home* (1898); HAYYIM SCHAUSS, *The Jewish Festivals* (1938; orig. pub. in Hebrew, 1933), and *The Lifetime of a Jew Throughout the Ages of Jewish History* (1950); ABRAHAM Z. IDELSOHN, *Jewish Liturgy and Its Development* (1932, reprinted 1967); MARK L. RAPHAEL (ed.), *Jews and Judaism in the United States: A Documentary History* (1983); SHALOM LILKER, *Kibbutz Judaism: A New Tradition in the Making* (1982).

Art and iconography: FRANZ LANDSBERGER, *A History of Jewish Art* (1946); ABRAHAM Z. IDELSOHN, *Jewish Music in Its Historical Development* (1929, reprinted 1967); CECIL ROTH (ed.), *Jewish Art: An Illustrated History* (1961).

Relation with non-Judaic religions: LEO BAECK, *Judaism and Christianity* (1958); SAMUEL SANDMEL, *We Jews and You Christians: An Inquiry into Attitudes* (1967); JACOB KATZ, *From Prejudice to Destruction: Anti-Semitism* (1980); FRANCIS E. PETERS, *Children of Abraham: Judaism, Christianity, Islam* (1982).

The role of Judaism in Western culture and civilization: CECIL ROTH, *The Jewish Contribution to Civilization* 3rd ed. (1956); DENNIS B. KLEIN, *Jewish Origins of the Psychoanalytic Movement* (1981).

The present-day forms of Judaism: LEON D. STITSKIN (ed.), *Studies in Torah Judaism* (1969); ALFRED JOSPE (ed.), *Tradition and Contemporary Experience: Essays on Jewish Thought and Life* (1970); ARNOLD JACOB WOLF (ed.), *Rediscovering Judaism: Reflections on a New Theology* (1965); *The Condition of Jewish Belief: A Symposium, Compiled by the Editors of Commentary Magazine* (1966); MORDECAI M. KAPLAN, *Judaism as a Civilization*, 2nd ed. (1957); SOLOMON POLL, *The Hasidic Community of Williamsburg: A Study in the Sociology of Religion* (1969); BERNARD MARTIN (ed.), *Contemporary Reform Jewish Thought* (1968); CHARLES LIEBMAN, "Orthodoxy in American Jewish Life," *American Jewish Yearbook* (1965); MORDECAI WAXMAN (ed.), *Tradition and Change: The Development of Conservative Judaism* (1958); JOSHUA ROTHENBERG, *The Jewish Religion in the Soviet Union* (1972).

The religious year: ROLAND DE VAUX, *op. cit.*, summarizes the contemporary state of biblical scholarship regarding the origin and development of the Jewish calendar, sabbath, and festivals. Other relevant works include: THEODOR H. GASTER, *Festivals of the Jewish Year* (1953), an anthropological, comparative, and often speculative approach to the sabbath and festivals; SHLOMO YOSEF ZEVIN, *ha-Moʿadim ba-halakhah* (1944), a modern classic (in Hebrew) treating talmudic and post-talmudic developments in the festival observances; and MENAHEM M. KASHER, *Torah Shelemah*, vol. 13 (1949), a comprehensive history of the Jewish calendar, also in Hebrew.

General introductions to Jewish philosophy: JULIUS GUTTMANN, *Die Philosophie des Judentums* (1933; Eng. trans., *Philosophies of Judaism*, 1964), the best general treatment of Jewish philosophy from the ancient to the modern period, ending with Franz Rosenzweig; NATHAN ROTENSTREICH, *Jewish Philosophy in Modern Times: From Mendelssohn to Rosenzweig* (1968), a good philosophical discussion of the major thinkers in modern Jewish philosophy, especially perceptive on Hermann Cohen. ALEXANDER ALTMANN, *Essays in Jewish Intellectual History* (1981), a collection of insightful scholarly essays on many aspects of Jewish philosophy.

Hellenistic philosophy: NAHUM GLATZER (ed.), *The Essential Philo* (1971), lengthy selections from the major works of Philo, with notes; HARRY A. WOLFSON, *Philo*, rev. ed., 2 vol. (1962), the most comprehensive study of Philo in any language, with special emphasis upon Philo's influence upon later philosophy.

Medieval philosophy: ISAAC HUSIK, *A History of Medieval Jewish Philosophy* (1940, reprinted 1969), a thorough examination of each of the major medieval Jewish philosophers from Isaac Israeli through Joseph Albo, with good bibliography; *Three Jewish Philosophers: Philo, Saadya Gaon, Jehuda Halevi*, trans. and ed. by HANS LEWY, ALEXANDER ALTMANN, and ISAAK HEINEMANN (1965), a good introductory anthology containing representative selections from these classical figures with perceptive introductions and explanatory notes; GEORGES VAJDA, *Introduction à la pensée juive du moyen âge* (1947), a good general survey organized around the major philosophical traditions in medieval Jewish philosophy, with extensive bibliography.

Jewish Kalām: SAʿADIA BEN JOSEPH, *The Book of Beliefs and Opinions*, trans. by SAMUEL ROSENBLATT (1948), Saʿadia's important philosophical work dealing with the major topics in Jewish theology such as faith and reason, creation, God, and reward and punishment; *A Karaite Anthology*, trans. and ed. by LEON NEMOY (1952), an excellent collection of the more accessible Karaite materials, covering a wide range of themes, writers, and periods; ISRAEL EFROS, "Medieval Jewish Philosophy" (1967), a collection of essays (in Hebrew) containing a full-length study of Saʿadia's philosophy, in addition to short essays on Judah ha-Levi and Maimonides.

Jewish Neo-Platonism: ISAAC ISRAELI, *Works*, ed. and trans. by ALEXANDER ALTMANN and S.M. STERN (1958), contains complete translations of three of Israeli's works and excerpts from another, together with a comprehensive essay on his philosophy; SOLOMON IBN GABIROL, "Fountain of Life," trans. from the Latin into Hebrew by J. BLUWSTEIN (1950), in addition to a complete vocalized translation of the text this edition contains the Hebrew summary made by Shem Tov ibn Falaquera in the 13th century.

Judah ha-Levi: *The Kuzari*, trans. by HARTWIG HIRSCHFELD (1964), a complete English translation of the text and notes, with an introductory essay by HENRY SLONIMSKY; LEO STRAUSS, "The Law of Reason in the Kuzari," in *Persecution and the Art of Writing* (1952), in addition to this perceptive and provocative essay on ha-Levi's attitude towards philosophy and rationalistic ethics, the book contains important essays on Maimonides and Spinoza; BAHYA IBN PAQUDA, *The Duties of the Heart* (Hebrew ed. by A. ZIFRONI, 1928; Eng. trans. by EDWIN COLLINS, 1904), one of the more widely read medieval classics of Jewish philosophy, concentrating upon ethics and personal piety; ABRAHAM BAR HIYYA, *The Meditation of the Sad Soul*, trans. with introduction by G. WIGODER (1971), one of the more accessible philosophical works of this important medieval astronomer and mathematician, devoted primarily to ethical and religious themes.

Jewish Aristotelianism: ABRAHAM IBN DAUD, *Das Buch Emunah Ramah*, ed. and trans. into German by S. WEIL (1852), one of the first manifestations of the Jewish assimilation of Aristotelian philosophy, containing a critique of Gabirol's neo-Platonism.

Maimonides: *The Guide of the Perplexed*, trans. with introductory essay by SHLOMO PINES (1963), an accurate and clear translation, with an excellent essay by Pines and a valuable prefatory essay by LEO STRAUSS; *A Maimonides Reader*, ed. with introduction and notes by I. TWERSKY (1972), a fine anthology containing important material from Maimonides' *Mishne Torah* and other legal writings, as well as from his shorter philosophical-theological essays and the *Guide*; HARRY A. WOLFSON, "Maimonides on Negative Attributes," in the *Louis Ginzberg Jubilee Volume*, pp. 411–446 (1945), the best historical and analytical study of Maimonides' doctrine of divine attributes; and "Halevi and Maimonides on Prophecy," *Jewish Quarterly Review*, n.s., 32:345–370, 33:49–82 (1942), an excellent historical analysis of the sources of doctrines of both philosophers on prophecy and a clear analysis of the differences between them. ISADORE TWERSKY, *Studies in Jewish Law and Philosophy* (1982), is a collection of previously published articles, some of them in Hebrew.

Averroists: (*Isaac Albalag*): GEORGES VAJDA, *Isaac Albalag, Averroïste juif, traducteur et annotateur d'Al-Ghazâlî* (1960), a translation of and commentary on Albalag's main philosophical work, al-Ghazālī's "Inconsistencies of the Philosophers," containing valuable citations from the classical Arabic philosophical texts. (*Levi ben Gerson*): "The Wars of the Lord" (1866), Gersonides' major philosophical work (in Hebrew) dealing with the most difficult and controversial topics in medieval philosophy and science; SEYMOUR FELDMAN, "Gersonides' Proofs for the Creation of the Universe," *Proceedings of the American Academy for Jewish Research*, 35:113–137 (1967), an analytical study of Gersonides' theory of creation and of his criticism of Aristotle's theory of eternity of the universe. (*Hasdai Crescas*): "The Light of the Lord" (1861), Crescas' most original critique (in Hebrew) of Aristotelian philosophy and his vigorous defense of traditional Judaism; HARRY A. WOLFSON, *Crescas' Critique of Aristotle* (1929), the most important study of medieval Jewish and Arabic philosophy so far written, containing a translation and critical Hebrew text of part 1 of the treatise with comprehensive, detailed, and most valuable notes and introductory essay; JOSEPH ALBO, *Book of Principles*, trans. with critical text by ISAAC HUSIK, 4 vol. (1946), a fine translation with helpful notes of Albo's treatise in Jewish dogmatics.

Modern Jewish philosophy: (*Iberian-Dutch philosophers*): I.S. RÉVAH, *Spinoza et le dr. Juan de Prado* (1959), a study of the cultural background of Spinoza's Amsterdam, especially of the heterodox elements in Sefardic Judaism, containing valuable material pertaining to the excommunication of Spinoza and the ideas of Uriel da Costa; BARUCH SPINOZA, *Theologico-Political Treatise*, trans. by R.H.M. ELWES (1883, reprinted 1951), Spinoza's critique of the Bible and the Jewish religion; LEO STRAUSS, *Spinoza's Critique of Religion* (1965), an excellent philosophical study of Spinoza's *Treatise*, its relation to Maimonides, Uriel da Costa, and other Sefardic heterodox thinkers; HARRY A. WOLFSON, *The*

Philosophy of Spinoza, 2 vol. (1934, reprinted 1969), a most detailed commentary on Spinoza's *Ethics*, containing valuable references to Spinoza's medieval sources such as Maimonides, Gersonides, and Crescas.

German philosophers: (*Moses Mendelssohn*): *Jerusalem and Other Jewish Writings*, trans. and ed. by ALFRED JOSPE (1969), a complete translation of *Jerusalem* and other miscellaneous writings, pertaining to questions on the Jewish religion, with a brief, informative, acute introduction by the editor; JULIUS GUTTMANN, "Mendelssohn's Jerusalem and Spinoza's *Theologico-Politico Treatise*" (in Hebrew) in his *Religion and Knowledge* (1956), a collection of Guttmann's essays that also includes important essays on ha-Levi, Maimonides, Gersonides, and Crescas, as well as on themes in modern philosophy of Judaism; NACHMAN KROCHMAL, "Collected Works" (1961, in Hebrew), containing his "Guide for the Perplexed of Our Times," ed. by S. RAWIDOWICZ and Rawidowicz' comprehensive analytical study of Krochmal's philosophy of Jewish history and its relationship to Hegelian philosophy. (*Hermann Cohen*): *Die Religion der Vernunft aus den Quellen des Judentums* (1919, reprinted 1966), Cohen's major work dealing with his philosophy of Judaism, organized according to the main themes of the Jewish religion; *Reason and Hope: Selections from the Jewish Writings of Hermann Cohen*, trans. and ed. by EVA JOSPE (1971), a useful collection from Cohen's miscellaneous writings on Jewish themes. (*Franz Rosenzweig*): *The Star of Redemption*, trans. by WILLIAM HALLO (1971), Rosenzweig's early but impressive statement of his "re-conversion" to Judaism, which has been influential in contemporary Jewish and non-Jewish theology; NAHUM GLATZER, *Franz Rosenzweig: His Life and Thought*, 2nd ed. (1961), an excellent anthology of Rosenzweig's various writings, containing many letters that reveal the more important and intimate episodes in Rosenzweig's career. (*Martin Buber*): See bibliography to article BUBER, MARTIN in the *Micropædia*. Interpretive studies include EMIL L. FACKENHEIM, *To Mend the World: Foundations of Future Jewish Thought* (1982); ARTHUR A. COHEN, *The Tremendum: A Theological Interpretation of the Holocaust* (1981); ALFRED JOSPE (ed.), *Studies in Jewish Thought: An Anthology of German Jewish Scholarship* (1981).

Jewish mysticism: G.G. SCHOLEM, *Major Trends in Jewish Mysticism*, rev. ed. (1961), the standard survey of the subject, with a chapter-by-chapter bibliography, *On the Kabbalah and its Symbolism* (Eng. trans. 1965), several studies on some of the great themes of Jewish mysticism; A.E. WAITE, *The Holy Kabbalah: A Study of Secret Tradition in Israel* (1929), a theosophical view of Jewish mysticism. HUGO ODEBERG, *3 Enoch; or the Hebrew Book of Enoch* (1928); ISIDOR KALISCH, *A Book on Creation* (1877); J. ABELSON, *The Immanence of God in Rabbinical Literature* (1912); G.G. SCHOLEM, *Jewish Gnosticism, Merkabah Mysticism, and Talmudic Tradition* (1960); JOSHUA TRACHTENBERG, *Jewish Magic and Superstition* (1939, paperback 1961). G.G. SCHOLEM, *Ursprung and Anfänge der Kabbala* (1962; French trans., *Les Origines de la Kabbale*, 1966); *Le commentaire d'Ezra de Gérone sur le Cantique des Cantiques*, trans. by G. VAJDA (1969); *The Zohar*, trans. by H. SPERLING and M. SIMON, 5 vol. (1931–34; paperback, sel. and ed. by G.G. SCHOLEM, 1963). MOSES CORDOVERO, *The Palm Tree of Deborah*, trans. from the Hebrew by L. JACOBS, 3rd ed. (1981); RAPHAEL J. ZWI WERBLOWSKY, *Joseph Karo: Lawyer and Mystic* (1962). L.I. NEWMAN, *The Hasidic Anthology* (Eng. trans. 1963); DOB BAER OF LUBAVITCH, *Tract on Ecstasy*, trans. from the Hebrew by L. JACOBS (1963). MARTIN BUBER, *The Origin and Meaning of Hasidism* (Eng. trans. 1960); S.H. DRESNER, *The Zaddik: The Doctrine of the Zaddik According to the Writings of Rabbi Yaakov Yosef of Polnoy* (1960); S. POLL, *The Hasidic Community of Williamsburg* (1962). J.L. BLAU, *The Christian Interpretation of the Cabala in the Renaissance* (1944); F. SECRET, *Les Kabbalistes chrétiens de la Renaissance* (1964); BEN ZION BOKSER, *The Jewish Mystical Tradition* (1981).

Jewish myth and legend: T.H. GASTER, *Myth, Legend, and Custom in the Old Testament* (1969); R.H. CHARLES (ed.), *The Apocrypha and Pseudepigrapha of the Old Testament in English*, 2 vol. (1913, reprinted 1964); M.R. JAMES, *The Lost Apocrypha of the Old Testament, Their Titles and Fragments* (1920). Myth and legend of the Hellenistic period is treated in W.N. STEARNS (ed.), *Fragments from Graeco-Jewish Writers* (1908); C.C. TORREY (ed. and trans.), *Lives of the Prophets* (1946); T.H. GASTER (trans.), *The Dead Sea Scriptures in English Translation*, pp. 256–267, 2nd ed., (1964). The Legends of the Talmud and Midrash are digested and annotated in LOUIS GINZBERG's classic, *Legends of the Jews*, 7 vol. (1909–39), also available in a one-volume abridgment (1961). The *Midrash Rabbah* has been translated and edited by HARRY FREEDMAN and MAURICE SIMON, 13 vol. in 5, 3rd ed. (1983); and the *Midrash on Psalms*, by W.G. BRAUDE, 2 vol. (1959). Principal editions and translators are listed in H.L. STRACK, *Einleitung in Talmud und Midraš*, 5th ed. (1921; Eng. trans., *Introduction to the Talmud and Midrash*, 1931). Compilations and studies of medieval myth and legend include *The Book of Jashar*, trans. by M.M. NOAH (1840); M.R. JAMES (trans.), *The Biblical Antiquities of Philo* (1917); MOSES GASTER (trans.), *The Chronicles of Jerahmeel* (1899); I.J. KAZIS (ed. and trans.), *The Book of the Gests of Alexander of Macedon* (in Hebrew; 1962); CURT LEVIANT (ed. and trans.), *King Artus* (1969); MOSES HADAS (trans.), *The Book of Delight* (1932). On the diffusion of medieval Jewish tales, see J. JACOBS, *Jewish Ideals, and Other Essays*, pp. 135–161 (1896); LOUIS GINZBERG, "Jewish Folklore: East and West," *Independence, Convergence, and Borrowing in Institutions, Thought, and Art*, Harvard Tercentenary Conference of Arts and Sciences, pp. 89–108 (1937). Judeo-German (Yiddish) works on the subject include MOSES GASTER, *Ma'aseh Book*, 2 vol. (1934); N.C. GORE (ed. and trans.), *Tzeenah u-Reenah: A Jewish Commentary on the Book of Exodus* (1965). Judeo-Spanish (Ladino) works include IGNACIO GONZALEZ LLUBERA (ed.), *Coplas de Yocef* (1935); C. CREWS, "Judaeo-Spanish Folktales in Macedonia," *Folk-Lore*, 43:193–225 (1932). For a treatment of Hasidic legend, see MARTIN BUBER (ed.), *Tales of the Hasidim*, 2 vol. (1947–48). Myths and legends of the Holy Land are treated in DOV NOY (ed.), *Folktales of Israel* (1963); ZEV VILNAY, *Legends of Palestine* (1932; orig. pub. in Hebrew, 1929), and his *The Sacred Land*, 3 vol. (1973–78).

Judicial and Arbitrational Systems

This article deals primarily with the operations of the judicial branch of government. It explores some of the fundamental relationships of this branch with legislative and executive branches and analyzes the functions, the structure and organization, and, finally, the key personnel of courts, such as judges and juries. This article also treats arbitration, another legal means of resolving disputes.

The approach is comparative, contrasting and comparing the systems of the two predominant legal traditions of the contemporary world: first, that of the common law, represented by England, the United States, Canada, Australia, and other nations deriving their legal systems from the English model; and, second, that of the civil law, as represented by nations of western Europe and Latin America and certain Asian and African nations that have modelled their legal systems on western European patterns. Reference is made to the legal institutions in the former Soviet Union and in eastern European nations. A separate section deals more specifically with judicial systems in Communist countries.

The article is divided into the following sections:

Functions of courts

KEEPING PEACE

The primary function of any court system in any nation—to help keep domestic peace—is so obvious that it is rarely considered or mentioned. If there were no agency to decide impartially and authoritatively whether a person had committed a crime and, if so, what should be done with him, other persons offended by his conduct would take the law into their own hands and proceed to punish him according to their uncontrolled discretion. If there were no agency empowered to decide private disputes impartially and authoritatively, self-help, quickly degenerating into physical violence, would prevail and anarchy would result. Not even a primitive society could survive under such conditions. All social order would be destroyed. In this most basic sense, courts constitute an essential element in society's machinery for keeping peace.

DECIDING CONTROVERSIES

In the course of helping to keep the peace, courts are called upon to decide controversies. If, in a criminal case, the defendant denies committing the acts charged against him, the court must choose between his version of the facts and the prosecution's; and if he asserts that his conduct did not constitute a crime, the court must decide whether his view of the law or the prosecution's is correct. In a civil case, if the defendant disputes the plaintiff's account of what happened between them—for example, whether they entered into a certain agreement—or if he disputes the plaintiff's view of the legal significance of whatever occurred—for example, whether the agreement was legally binding—the court again must choose between the contentions of the parties. The issues presented to, and decided by, the court may be either factual, legal, or both.

It would be a mistake, however, to assume that courts spend all of their time deciding controversies. Many cases brought before them are not contested. They represent potential, rather than actual, controversies in which the court's role is more administrative than adjudicatory. The mere existence of a court renders unnecessary any very frequent exercise of its powers. The fact that it operates by known rules and with reasonably predictable results leads those who might otherwise engage in controversy to compose their differences.

The court's administrative role

Most people arrested and charged with crime in the common-law world plead guilty. If they do so understandingly and without coercion of any sort, there is no need to determine guilt, for the sole question is whether the defendant should go to jail, pay a fine, or be subjected to other corrective treatment. In civil-law countries some judicial inquiry into the question of guilt or innocence is required even after a confession. But the inquiry is brief and tends to be perfunctory. The main problem to be resolved, usually without contest, is what sentence should be imposed.

The vast majority of civil cases are also uncontested or, at least, are settled before trial. The court keeps the calendar moving, sometimes encouraging settlement, and decides such questions of law or fact as are presented by the parties; but the number of cases actually tried is small compared to the number settled.

Most divorce cases are uncontested, both parties usually being anxious to terminate the marriage and often agreeing on related questions concerning support and the custody of children. All the court does in such cases is to review what the parties have agreed upon and give its official approval.

Many other uncontested matters come before courts, such as the adoption of children, the distribution of assets in trusts and estates, and the setting up of corporations. Occasionally questions of law or fact arise that have to be decided by the court, but normally all that is required is judicial supervision and approval.

JUDICIAL LAWMAKING

As courts decide controversies they create an important by-product beyond the peaceful settlement of disputes, that is, the development of rules for future cases. Law is thus made not only by legislatures but also by the courts.

To an extent that varies greatly between common-law and civil-law nations, all courts apply preexisting rules formulated by legislative bodies. In the course of doing so, they interpret those rules, sometimes distorting them, sometimes transforming them from generalities to specifics, sometimes filling gaps to cover situations never considered by the original lawmakers. The judicial decisions embodying these interpretations then become controlling for future cases, sometimes to the extent of virtually supplanting the legislative enactments themselves.

The uses of experience in the law: stare decisis

This is one aspect of the doctrine of precedent, or, as it is sometimes called, stare decisis (literally, "to stand by decided matters"). Judges follow earlier decisions, not only to save themselves the effort of working out fresh solutions for the same problems each time they recur but also, and primarily, because their goal is to render uniform and stable justice. If one individual is dealt with in a certain way today, the theory is that another individual engaging in substantially identical conduct under substantially identical conditions tomorrow or a month or year hence should be dealt with in the same way. This, reduced to its essentials, is all that precedent means.

In civil-law nations all judicial decisions are, in theory, based upon legislative enactments, and the doctrine of judicial precedent does not apply. Practice, however, departs from theory. While there are comprehensive legislative codes in these countries, supposedly covering almost every aspect of human conduct and supplying ready-made answers for all problems that can arise, in fact many of the provisions are exceedingly vague and are sometimes almost meaningless until applied to concrete situations, when judicial interpretation gives them specific meaning. Furthermore, the legislative codes cannot anticipate all situations that may arise and come before the courts. The gaps in legislative schemes must be and are filled by judicial decisions, for no court in any nation is likely to refuse to decide a case on the ground that it has not been told in advance the answers to the questions presented to it. Decisions dealing with circumstances unforeseen by the codes and giving specific meaning to vague legislative provisions are published in most civil-law countries and are frequently referred to by lawyers and relied upon by judges. They are not considered "binding," but neither are they forgotten or disregarded. In actual practice, they have almost as much influence as statutory interpretations in nations that formally adhere to the doctrine of stare decisis.

It remains true that in common-law countries judicial lawmaking is more pervasive and more frankly acknowledged than in civil-law countries. In addition to rendering decisions that authoritatively interpret statutes, the courts of these nations have created a vast body of law without any statutory foundation whatever. Centuries ago, when there was no legislation to guide them, judges began to decide cases in accordance with their own conceptions of justice. Later judges followed them, deciding like cases in the same manner but distinguishing earlier cases when dissimilar factors were discovered in the cases before them. The later cases also became precedents to be followed in still later cases presenting substantially similar fact patterns. So the process has continued over centuries and is still continuing. The total accumulation of all these judicial decisions is what constitutes "the common law"—the by-product of judges deciding cases and setting forth their reasons. In the common-law nations, legislation is, as a result, more limited in scope than in the civil-law countries. It does not purport to provide for all possibilities but leaves large areas of conduct to be governed solely by judge-made law.

Precedent and common law

To speak of precedent as "binding" even in common-law systems is misleading. As already noted, earlier decisions can be and are distinguished when judges conclude that they are based upon situations different from those before the court in later cases. Even more significant, earlier decisions can be overruled by the courts that rendered them (not by courts lower in the judicial hierarchy) when the judges conclude that they have proved to be so erroneous or unwise as to be unsuited for current or future application. The Supreme Court of the United States has overruled many of its own earlier decisions, to the consternation of those who yearn for a rigid separation of powers and who are unable to accept the inevitability of judicial lawmaking. Many of these overrulings are in the field of constitutional law, in which legislative correction of an erroneous judicial interpretation of the Constitution is impossible and in which the only alternative is the exceedingly slow, cumbersome, costly, and difficult process of constitutional amendment. Nevertheless, the power to overrule decisions is not restricted to constitutional interpretations. It extends to areas of purely statutory and purely judge-made law as well, areas in which legislative action would be equally capable of accomplishing needed changes. Even in England, which has no written constitution and which has traditionally followed a far more rigid doctrine of stare decisis than the United States, the House of Lords, in its role as the highest court, has announced its intention of departing from precedent "in appropriate cases."

Conflicting views of the court's role

The desirability of judicial lawmaking has long been the subject of lively debate in both civil- and common-law countries. That courts should not arrogate to themselves unrestricted legislative power is universally accepted. But when existing statutes and precedents are outmoded or barbarous as applied to specific cases before the courts, should not judges be able to change the law in order to achieve what they conceive to be just results or, stated differently, to avoid what they consider unjust results?

The extent to which the judges should be bound by statutes and case precedents as against their own ethical ideas and concepts of social, political, and economic policy is an important question, as is the matter of which should prevail when justice and law appear to the judges to be out of alignment with each other. These are questions upon which reasonable persons disagree vigorously even when they are in basic agreement on the proposition that some degree of judicial lawmaking is inevitable. What is mainly at issue is the proper tempo and scope of judicial change. How quickly should judges act to remedy injustice and when should they consider an existing rule to be so established that its alteration calls for constitutional amendment or legislative enactment rather than judicial decision? As many dissenting opinions attest, judges themselves disagree on the answers to these questions, even when they are sitting on the same bench hearing the same case.

CONSTITUTIONAL DECISIONS

Marbury v. *Madison*

In some nations courts not only interpret legislation but determine its validity and in so doing sometimes render statutes inoperative. This happens only in nations that have written constitutions and have developed a doctrine of "judicial supremacy." The prime example is the United States, and the classic statement of the doctrine is the Supreme Court's decision in *Marbury* v. *Madison* (1803), in which Chief Justice Marshall said:

> The powers of the legislature are defined and limited; and that those limits may not be mistaken, or forgotten, the Constitution is written. To what purpose are powers limited, and to what purpose is that limitation committed to writing, if these limits may, at any time, be passed by those intended to be restrained? The distinction between a government with limited and unlimited powers, is abolished, if those limits do not confine the persons on whom they are imposed, and if acts prohibited and acts allowed, are of equal obligation. It is a proposition too plain to be contested, that the Constitution controls any legislative act repugnant to it. . . . It is emphatically the province and duty of the judicial department to say what the law is. Those who apply the rule to particular cases, must of necessity expound and interpret that rule. If two laws conflict with each other, the courts must decide on the operation of each.

Armed with the authority asserted at this early date, the Supreme Court of the United States has held many statutes, federal as well as state, unconstitutional and has also invalidated executive actions that violated the Constitution. Even more surprising is the fact that lower courts also possess and exercise the same powers. Whenever a question arises in any U.S. court at any level as to the constitutionality of a statute or executive action, that court is obligated to determine its validity in the course of deciding the case before it. The case may have been brought for the sole and express purpose of testing the constitutionality of the statute or it may be an ordinary civil or criminal case, in which a constitutional question incidental to the main purpose of the proceeding is raised. Of course, when a lower court decides a constitutional question, its decision is subject to appellate review, sometimes at more than one level. When a state statute is challenged as violating the state constitution, the final authority is the supreme court of that state; when a federal or state statute or a state constitutional provision is challenged as violating the Constitution of the United States, the ultimate arbiter is the Supreme Court of the United States.

In a few American states, questions as to the constitutional validity of a statute may be referred in abstract form to the state's highest court by the chief executive or the legislature for an advisory opinion. This, however, is unusual and, in any event, supplementary to the normal procedure of raising and deciding constitutional questions. The normal pattern is for a constitutional question to be raised at the trial-court level in the context of a genuine controversy and to be decided finally on appellate review of the trial-court decision.

Other methods of constitutional adjudication

The U.S. pattern of constitutional adjudication is not followed in all nations that have written constitutions. In some, such as Germany, there is a special court at the highest level of government that handles only constitutional questions and to which all such questions are referred as soon as they arise. A constitutional question may be referred to the special court in abstract form for a declaratory opinion by a procedure similar to that prevailing in the minority of U.S. states that allow advisory opinions.

In other nations, written constitutions may be in effect but not accompanied by any conception that their authoritative interpretation is a judicial function. Legislative bodies, rather than courts, act as the guardians and interpreters of the constitution, being guided by their provisions but not bound by them in any realistic sense.

Finally, there are some nations, such as England, that have no written constitutions. Here parliamentary supremacy clearly prevails. The courts have no power to invalidate statutes, although they can and do interpret them.

PROCEDURAL RULE MAKING

Distinct from the type of lawmaking just described is a more conscious and explicit type of judicial legislation and one that is less controversial. It is directed toward the rules of procedure by which the courts operate. This is a technical area in which expert knowledge of the type possessed by judges and lawyers is needed; in which constant attention to detail is required; and in which major problems of social, economic, or political policy are seldom encountered. Some legislative bodies, able or willing to devote only sporadic attention to the day-to-day problems of the management of litigation, have delegated the power to regulate procedure to the courts themselves. This is not ad hoc judicial lawmaking as a by-product of deciding cases but openly acknowledged promulgation of general rules for the future, in legislative form, by courts rather than legislatures.

An outstanding example of judicial rule making is found in the United States, where Congress has delegated to the Supreme Court broad power to formulate rules of civil, criminal, and appellate procedure for the federal courts. The Supreme Court also has and exercises the power to amend the rules from time to time as experience indicates that changes are desirable. Congress reserves the power to veto the rules so promulgated but has felt no need to exercise it.

Other legislative bodies, including those of some American states and most of the nations of continental Europe, have been unwilling to repose equal trust in the courts and have retained for themselves the power to regulate procedure. The results have been varied. Courts sometimes become so immersed in day-to-day decision making that they fail to pay adequate attention to the proper functioning of the judicial machinery and so perpetuate rules that are unduly rigid, unrealistic, and unsuited to the needs of litigants, which was the case in England and the American colonies during the 18th and first part of the 19th century. When such a condition occurs, reform through legislative action is indicated. Apart from the occasional necessity of major sweeping changes, however, experience in the common-law countries, at least, indicates that procedural rule making is better vested in the courts than in legislative bodies.

REVIEW OF ADMINISTRATIVE DECISIONS

Existing alongside the courts in any nation are administrative agencies of various kinds. Some do substantially the same kind of work as is done by courts and in substantially the same manner; some have quite different functions such as the issuing of licenses and the payment of welfare benefits.

The relationship between such agencies and regular courts differs markedly between common-law countries and civil-law countries. In common-law countries the actions of administrative agencies are subject to review in the ordinary courts. If the agency is one that decides controversies in substantially the same manner as a court, but in a different and more limited area, judicial control takes much the same form of appellate review as is provided for the decisions of lower courts. The objective of reviewing the record of proceedings is to determine whether the administrative agency acted within the scope of its jurisdiction, whether there was any evidence to support its conclusion, and whether the governing law was correctly interpreted and applied. Administrative decisions are seldom upset by the courts because of a belief on the part of most judges that administrative agencies have special expertise in the area of their specialty. However, they can be and occasionally are upset, thus underscoring the large degree of judicial control over other agencies of government that characterizes common-law systems. If the administrative agency does not engage in formal adjudication, it produces no record of proceedings for judicial review. Nevertheless, its action can be challenged in court by way of trial rather than appeal. The same problems are presented for judicial determination: did the agency act within its jurisdiction, did it correctly follow the law, and was there any rational or factual basis for its action?

In many civil-law countries, the ordinary courts have no control over administrative agencies. Their decisions are reviewed by a special tribunal that is engaged exclusively in that work and that has nothing to do with cases of the type that come into the courts. Its function is solely appellate and solely within the specialized areas entrusted to the administrative agencies. The prototype of this type of tribunal is the Conseil d'État of France.

ENFORCEMENT OF JUDICIAL DECISIONS

The method of enforcing a judicial decision depends upon its nature. If it does nothing more than declare legal rights, as is true of a simple divorce decree (merely severing marital ties, not awarding alimony or the custody of children), or a declaratory judgment (for example, interpreting a contract or a statute), no enforcement is needed. If a judgment orders a party to do or refrain from doing a certain act, as happens when an injunction is issued, the court itself takes the first step in enforcing the judgment by holding in contempt anyone who refuses to obey its order and sentencing him to pay a fine or go to jail. Thereafter, enforcement is in the hands of the executive branch of government, acting through its correctional authorities.

In routine criminal cases and in civil cases that result in the award of money damages, courts have little to do with the enforcement of their judgments. That is the function of the executive branch of government, acting through sheriffs, marshals, jailers, and similar officials. The courts themselves have no machinery for enforcement.

The limits of judicial power

Some judgments are extremely controversial, as was the case with the decision of the Supreme Court of the United States ordering racial desegregation of the schools. When voluntary compliance with such a judgment is refused, forcible methods of enforcement are necessary, sometimes extending to the deployment of armed forces under the control of the executive branch of the government. The withdrawal of executive support seldom occurs, even when decisions are directed against the executive branch itself; but when such executive support is withheld, the courts are rendered impotent. Judges, being aware of their limited power, seldom render decisions that they know to be so lacking in support that they will not be enforced.

Court structure and organization

TYPES OF COURTS

There are many different types of courts and many ways to classify and describe them. Basic distinctions must be made between civil and criminal courts, between courts of general jurisdiction and those of limited jurisdiction, and between trial and appellate courts.

Criminal courts. Criminal courts deal with persons accused of crime, deciding whether they are guilty and, if so, determining the consequences they shall suffer. Prosecution is on behalf of the public, represented by some official such as a district attorney, procurator, or a police officer. Courts are also public agencies, but in this instance they stand neutral between the prosecution and the defense, their objective being to decide between the two in accordance with law.

In civil-law countries a more active role is assigned to the judge and a more passive role to counsel than in common-law countries. In the common-law courts, in which the "adversary" procedure prevails, the lawyers for both sides bear responsibility for producing evidence and they do most of the questioning of witnesses. In civil-law countries, "inquisitorial" procedure prevails, with judges doing most of the questioning of witnesses and having an independent responsibility to discover the facts. This difference pertains more to procedure rather than function.

Purposes of sentences

If a person has been found guilty, he is sentenced, again according to law and within limits fixed by legislation. The objective is not so much to wreak vengeance upon the offender as to rehabilitate him and deter others from following his example. Hence the most common sentences are fines, short terms of imprisonment, and probation (which allows the offender to remain at large but under supervision). In extremely serious cases, the goal may be to prevent the offender from committing further crimes, which may call for a long term of imprisonment or even capital punishment. The death penalty, however, is gradually disappearing from the criminal codes of civilized nations.

Criminal proceedings in any nation inevitably have some educational impact on defendants and on members of the general public. In Communist nations education is a conscious and primary goal. A basic provision of old Soviet law declared:

> By all its activities the court shall educate the citizens of the U.S.S.R. in the spirit of devotion to the Motherland and the cause of communism in the spirit of strict and undeviating observance of Soviet laws, of care for socialist property, of labor discipline, of honesty toward public and social duty, of respect for the rights, honor and dignity of citizens, for the rules of socialist common-life.

Civil courts. Civil courts deal with "private" controversies, as where two individuals (or corporations) are in dispute over the terms of a contract or over who shall bear responsibility for an auto accident. Ordinarily the public is not a party as in criminal proceedings, for it has no interest beyond providing just rules for decision and a forum where the dispute can be impartially and peacefully resolved.

It is possible, however, for the government to be involved in civil litigation if it stands in the same relation to a private party as another individual might stand. Thus, if a postal truck should run down a pedestrian, the government might be sued civilly by the injured person; or if the government contracted to purchase supplies that turned out to be defective, it might sue the dealer for damages in a civil court.

The objective of a civil action is not punishment or correction of the defendant or the setting of an example to others but rather to restore the parties so far as possible to the positions they would have occupied had no legal wrong been committed. The most common civil remedy is a judgment for money damages, but there are others, such as an injunction ordering the defendant to do, or refrain from doing, a certain act or a judgment restoring property to its rightful owner.

Civil claims do not ordinarily arise out of criminal acts. A person who breaks his contract with another or who causes him a physical injury through negligence may have committed no crime but only a civil wrong for which he may not be prosecuted criminally by the public.

There are, however, areas of overlap, for a single incident may give rise to both civil liability and criminal prosecution. In some nations, such as France, both types of responsibility can be determined in a single proceeding under a concept known as adhesion by which the injured party is allowed to assert his civil claim in the criminal prosecution, agreeing to abide by its outcome. This removes the necessity of two separate trials. In common-law countries there is no such procedure, even though civil and criminal jurisdiction may be merged in a single court. Two separate actions must be brought, independent of each other.

Courts of general jurisdiction. Although there are some courts that handle only criminal cases and others that handle only civil cases, a more common pattern is for a single court to be vested with both civil and criminal jurisdiction. Such is the High Court of England and such are many of the trial courts found in U.S. states. Often these tribunals are called courts of general jurisdiction, signifying that they can deal with almost any type of controversy, although in fact they may not have jurisdiction over certain types of cases assigned to specialized tribunals. Often such courts are also described as superior courts, because they are empowered to handle serious criminal cases and important civil cases involving large amounts of money.

Even if a court possesses general or very broad jurisdiction, it may nevertheless be organized into specialized branches, one handling criminal cases, another handling civil cases, another handling juvenile cases, and so forth. The advantage of such an arrangement is that judges can be transferred from one type of work to another, and cases do not fail to be heard for having been instituted in the wrong branch since they can be transferred administratively with relatively little effort.

Specialized tribunals

Courts of limited jurisdiction. Specialized tribunals of many kinds exist, varying from nation to nation. Some deal only with the administration of the estates of deceased persons (probate courts), some only with disputes between merchants (commercial courts), some only with disputes between employers and employees (labour courts). All are courts of limited jurisdiction. Deserving of special mention because of their importance are juvenile courts, empowered to deal with misconduct by children and sometimes also with the neglect or maltreatment of children. Their procedure is much more informal than that of adult criminal courts, and the facilities available to them for the pretrial detention of children and for their incarceration, if necessary after trial, are different. The emphasis is on salvaging children, not punishing them.

Traffic courts also deserve mention because they are so common. They process motor vehicle offenses such as speeding and improper parking. Their procedure is summary and their volume of cases heavy. Contested trials are relatively infrequent.

Finally, in most jurisdictions there are what are called, unfortunately and for want of a better term, "inferior" courts. These are often manned by part-time judges who are not trained in the law. They handle minor civil cases involving small sums of money, such as bill collections, and minor criminal cases carrying light penalties, such as simple assaults. In addition to finally disposing of minor criminal cases, such courts may handle the early phases of more serious criminal cases—fixing bail, advising defendants of their rights, appointing counsel, and conducting preliminary hearings to determine whether the evidence is sufficient to justify holding defendants for trial in higher "superior" courts.

Appellate courts. The tribunals described thus far are trial courts or "courts of first instance." They see the parties, hear the witnesses, receive the evidence, find the facts, apply the law, and determine the outcome.

Above them, to review their work and correct their errors, are appellate courts. These are usually collegiate bodies, consisting of several judges instead of the single judge who usually presides over a trial court. The jurisdiction of the appellate courts is usually general; specialized appellate tribunals handling, for example, only criminal appeals or only civil appeals are rare, although not unknown. The Conseil d'État of France and the Federal Constitutional Court of Germany have already been mentioned as examples of specialization.

Appellate review is not automatic. It must be sought by some party aggrieved by the judgment in the court below. For that reason, and because an appeal may be both expensive and useless, there are far fewer appeals than trials and, if successive appeals are available, as is often the case, far fewer second appeals than original appeals. Judicial systems are organized on a hierarchical basis: at the bottom are numerous trial courts scattered throughout the nation; above them are a smaller number of first-level appellate courts, usually organized on a regional basis; and at the apex is a single court of last resort.

Varieties of appellate review

There are three basic types of appellate review. The first consists of a retrial of the case, with the appellate court hearing the evidence for the second time, making fresh findings of fact, and in general proceeding in much the same manner as the court that originally rendered the judgment under attack. This "trial de novo" is used in common-law countries for the first stage of review but only when the trial in the first instance was conducted by an "inferior" court—one typically manned by a part-time judge or two or more such judges, empowered to try only minor cases and keeping no adequate record of its proceedings.

The second type of review is based in part on a "dossier," which is a record compiled in the court below of the evidence received and the findings made there. The reviewing court has the power to rehear the same witnesses again or to supplement their testimony by taking additional evidence, but it need not and frequently does not do so, being content to rely on the record already made in reaching its own findings of fact and conclusions of law. This type of proceeding prevails generally in civil-law countries for the first stage of appellate review, even when the original trial was conducted in a superior court, staffed by professional judges, and empowered to try important or serious cases.

The third type of review is based solely on a written record of proceedings in the court or courts below. The reviewing court does not itself receive evidence directly but concentrates its effort on discovering from the record whether any errors were committed of such a serious nature as to require reversal or modification of the judgment under attack or a new trial in the court below. The emphasis is on questions of law (both procedural and substantive) rather than on questions of fact. This type of review prevails both in civil-law nations and common-law nations at the highest appellate level. It is also used in common-law nations at lower levels when the judgment of a superior court is under attack. The purpose of this type of review is not merely to assure that correct results are reached in individual cases but also to clarify and expound the law in the manner described earlier. Lower courts have little to do with the development of the law, for they ordinarily do not write or publish opinions. The highest appellate courts do, and it is their opinions that become the guidelines for future cases.

Courts in federal systems. Many nations, such as England, France, and Japan, have unitary judicial systems with all courts (that is, regular courts as distinguished from administrative bodies) fitting into a single national hierarchy of tribunals along the lines just described. Other nations, organized on a federal basis, tend to have more complicated court structures, reflecting the fragmentation of governmental powers between the central authority and the local authorities. In the United States, for example, there are 51 separate judicial systems, one for each state and another for the federal government. To a limited extent, the jurisdiction of the federal courts is exclusive of that exercised by the state courts, but there are large areas of overlap and duplication. At the top level is the Supreme Court of the United States, hearing appeals not only from the lower federal courts but also from state courts insofar as they present federal questions arising under the Constitution of the United States or under federal statutes or treaties. If a case in a state court involves only a question of state law—for example, the interpretation of a state statute—the ultimate authority is the state supreme court, and no appeal is possible to the Supreme Court of the United States.

Court structure in a federal form of government need not be as complicated as that in the United States. It is possible to have only one set of courts for the nation, operated by the central government and handling all cases that arise under state law as well as federal law.

Another possibility is for each state or province to have its own system of courts, handling all questions of federal as well as state law, and for the central government to maintain only a single supreme court to decide questions as to the relationship of the central authority and the local authorities or as to the relationship among the local authorities themselves. This is the pattern in Canada and Australia.

Conflict of laws problems

Another complication resulting from a federal form of government is that questions involving conflict of laws arise with great frequency. Such questions concern the choice to be made between the law of one jurisdiction and another as the rule for decision in a particular case. Even in a unitary system, such problems cannot be avoided, for an English court may be called upon to try a case arising from a transaction that took place in France and to decide whether English or French law should govern. Such problems arise much more often, however, in federal systems, where laws differ from state to state and people move about very freely. Their activities in one state sometimes become the subject of a lawsuit in another, requiring the court to decide which law should apply.

JUDGES

A court is a complex institution whose functioning depends upon many people: not only the judge but also the parties, their lawyers, witnesses, clerks, bailiffs, probation officers, administrators, and many others, including, in certain types of cases, jurors. Nevertheless, the central figure in any court is the judge.

Judges vary enormously, not only from nation to nation but often within a single nation. For example, a rural justice of the peace in the United States—untrained in the law, serving part-time, sitting alone in work clothes in a makeshift courtroom, collecting small fees or receiving a pittance for salary, trying a succession of routine traffic cases and little else—obviously bears little resemblance to a justice of the Supreme Court of the United States—a full-time, well-paid, black-robed professional, assisted by law clerks and secretaries, sitting in a marble palace with eight colleagues and deciding at the highest appellate level only questions of profound national importance. Yet both persons are judges.

The levels of legal training

Lay judges. In some civil-law countries, judges at all levels are professionally trained in the law, but in many other nations they are not. In England, part-time lay judges outnumber full-time professional judges by about 60 to 1. Called magistrates or justices of the peace, they dispose of about 97 percent of all criminal cases in that nation and do so with general public satisfaction and the approbation of most lawyers. Professional judges deal only with the most serious crimes, which are relatively few in number; most of their time is devoted to civil cases. England places unusually heavy reliance on lay judges, but they are far from unknown in the courts of many other nations, particularly at the lowest trial level. This was also true in the former U.S.S.R. and remains so in the United States. There is considerable diversity in the way laymen are chosen and used in judicial work. In the United States, for example, lay judges are popularly elected for limited terms, whereas in England they are appointed by the lord chancellor to serve until retirement or removal. In England the lay judges serve intermittently in panels on a rotating basis for short periods, whereas in the United States they sit alone and continuously. In the erstwhile U.S.S.R. lay judges (called assessors) always sat with professional judges; in England, they sometimes do; and in the United States, they never do. In some underdeveloped nations, few judges at any level are legally trained. They are more likely to be priests, for the law they administer is mainly derived from religious teaching, and religion and secular government are often not sharply differentiated. The vast majority of nations that use lay judges at the lowest trial level, however, insist upon professionally trained judges at higher levels: in trial courts of general jurisdiction and in appellate courts.

Professional judges in the civil-law tradition. Professional judges in civil-law countries are markedly different in background and outlook from professional judges in common-law countries. Both are law-trained and both perform substantially the same functions, but there the similarities cease. In a typical civil-law country, a person graduating from law school makes a choice between a judicial career and a career as a private lawyer. If he chooses the former and is able to pass an examination, he is appointed to the judiciary by the minister of justice (a political officer) and enters service in his early 20s. His first assignment is to a low-level court; thereafter, he works his way up the judicial ladder as far as he can until his retirement on a pension. His promotions and assignments depend upon the way his performance is regarded by a council of senior judges, or sometimes upon the judgment of the minister of justice, who may or may not exercise his powers disinterestedly and on the basis of merit. The civil-law judge, in short, is a civil servant.

Professional judges in the common-law tradition. In common-law nations, the path to judicial office is quite different. Upon completion of his formal education, a person typically spends 15, 20, or 25 years in the private practice of law or, less commonly, in law teaching or governmental legal service and then, at about age 50, becomes a judge. He takes no competitive examination but is appointed or elected to office. In England the appointive system prevails for all levels of judges, including even lay magistrates. Appointments are primarily under the control of the lord chancellor, who, although a cabinet officer, is also the highest judge of the realm. They are kept surprisingly free from party politics. In the United States, the appointive method is used in federal courts and some state courts, but it tends to be highly political. Appointments are made by the chief executive of the nation or state and are frequently subject to legislative approval. In many states, judges are popularly elected, sometimes on nonpartisan ballots, sometimes on partisan ballots with all the trappings of traditional political contests. In an attempt to de-emphasize political considerations and yet maintain some measure of popular control over the selection of judges, a third method of judicial selection has been devised and is slowly growing in popularity. Called the Missouri Plan, it involves the creation of a nominating commission that screens judicial candidates and submits to the appointing authority a limited number of names of persons considered qualified. The appointing authority must make his choice from the list submitted. The person chosen as judge then assumes office for a limited time, and, after the conclusion of this probationary period, he stands for "election" for a much longer term. He does not run against any other candidate but only "against his own record."

Political status of the judiciary

In common-law countries, a person does not necessarily enter the judiciary at a low level; he may be appointed or elected to his nation's highest court or to one of its intermediate courts. He does not look forward to any regular pattern of promotion, nor is he necessarily assured of long tenure with ultimate retirement on a pension. In some courts, life tenure is provided, usually subject to mandatory retirement at a fixed age. In others, tenure is limited to a stated term of years. At the conclusion of his term, if not mandatorily retired earlier, the judge must be reelected or reappointed if he is to continue.

While in office, the common-law judge enjoys greater power and prestige and more independence than his civil-law counterpart. He occupies a position to which most members of his profession aspire. He is not subject to outside supervision and inspection by any council of judges or by a minister of justice; nor is he liable to be transferred by action of such an official from court to court or place to place. The only administrative control over him is that exercised by judicial colleagues, whose powers of management are generally slight, being limited to such matters as requiring periodical reports of pending cases and arranging for temporary (and usually consensual) transfers of judges between courts when factors such as illness or congested calendars require them. Only if a judge misbehaves very badly is he in danger of disciplinary sanctions and then usually only by way of criminal prosecution for his misdeeds or legislative impeachment and trial, resulting in removal from office—a very cumbersome, slow, ill-defined, inflexible, ineffective, and seldom used procedure. In parts of the United States, newer and more expeditious methods of judicial discipline are developing in which senior judges are vested with power to impose sanctions ranging from reprimand to removal from office of erring colleagues. They are also vested with power to retire judges who have become physically or mentally unfit to discharge their duties.

Discipline of judges

Except at the very highest appellate level, common-law judges are no less subject than their civil-law counterparts to appellate reversals of their judgments. But appellate review cannot fairly be regarded as discipline. It is designed to protect the rights of litigants; to clarify, expound, and develop the law; and to help and guide rather than reprimand lower court judges. (D.K.)

JURIES

The jury is a historic legal institution in which a group of laymen participate in a major way in deciding cases brought to trial. Its exact characteristics and powers depend on the laws and practices of the countries, provinces, or states in which it is found, and there is considerable variation. Basically, however, it recruits laymen at random from the widest population for the trial of a particular case and allows them to deliberate in secrecy, to reach a decision by other than majority vote, and to make it public without giving reasons.

History and use. The jury's origin is lost in the past. It may have been indigenous to England or have been brought there by the Norman invaders in 1066. Originally, the jurors were neighbourhood witnesses who passed judgment based on what they themselves knew. But the breakdown of medieval society and the growth of the towns changed this; the jury was called upon to determine the facts of the case, based upon the evidence presented in court. The availability of the jury in the king's courts may have been a key factor in centralizing the nation's courts under the king and in creating the common law. By the 15th century, nonrational modes of trial such as ordeal, in which the defendant was subjected to various tortures that, if successfully endured, proved his innocence, were replaced by the jury trial, which became the established form of trial for both criminal and civil cases at common law.

Two forces moved the jury abroad. One was the expansion of the British Empire, which brought the jury to Asia, Africa, and the American continent. The other was the French Revolution and its aftermath, which brought it, as a symbol of popular government, to the European continent: first to France itself, then, through Napoleon, to the Rhineland, later to Belgium, most of the remaining German states, Austria-Hungary, Russia, Italy, Switzerland, Holland, and Luxembourg, although the last two abolished it immediately after Napoleon's defeat. In each of these countries, use of the jury was from the outset limited to trials of major crimes and of political crimes against the state.

Beginning in the mid-19th century, the jury was weakened in a variety of ways: in 1850, Prussia, for example, removed treason from its jurisdiction; in 1851, the duchy of Nassau removed all political crimes; in 1923, Czechoslovakia removed treason and, one year later, libel; in 1919, Hungary suspended jury trial entirely and never restored it. Germany abandoned the jury in 1924. Both the Soviet bloc and the fascist states abolished it outright; France never restored the jury abolished during the German occupation in the 1940s, and Japan did away with its short-lived jury courts in 1943. After World War II, Austria reintroduced the jury in a weakened form.

United States as major user of jury system

Thus, there are three important points about the history and development of the jury as a legal institution: first, the effort to introduce it outside the Anglo-American legal orbit has failed; further, in England itself its use was limited by statute to a small category of cases; and, thus, the United States has emerged today as the home of the jury system for both criminal and civil cases. Some 120,000 jury trials are conducted there annually, more than 90 percent of all jury trials in the world.

Use of the jury in the United States depends on two factors: the degree to which it is available as a matter of right and the degree to which the parties themselves choose to use it. The laws as to its availability have varied from state to state, but in 1968 in *Duncan* v. *Louisiana* the United States Supreme Court declared that a jury trial is a constitutional right in all criminal cases in which the penalty may exceed six months' imprisonment. In civil cases its constitutional status is less clear, but, in general, jury trial is available. The practice of allowing the parties to waive a jury trial also varies widely from region to region, and, as a result, the number of jury trials per year also varies widely. The annual number of criminal jury trials per 100,000 population ranges between 3 for Connecticut to 144 for Georgia.

Jury procedures. *Selection.* Historically there were some minimum requirements of property and competence for jury service. More recently the idea of genuine random selection from the population, to achieve a cross section of the community, has been gaining ground. Since 1969 it has been the principle of selection in the federal courts. Most jurisdictions exempt some groups from jury service: police officers, lawyers, doctors, and so on. All jurisdictions excuse jurors if the service imposes undue hardship.

The voir dire

The commitment of important decisions to a random group of laypersons has been moderated, particularly in the United States, by an elaborate screening, voir dire, conducted by trial counsel at the inception of a trial. The law permits counsel to challenge prospective jurors either for cause (if there is specific likelihood of bias) or, for a limited number, "peremptorily"—that is, without having to give a reason. American trial tradition attaches a great deal of significance to the strategies of juror selection, and in celebrated cases the lawyers' voir dire examination has extended for several weeks.

Size and unanimity. Traditionally the jury had 12 members and was required to reach its decisions with unanimity, a striking arrangement in Anglo-American countries that make all other decisions by majority vote. Over the years some modifications have been made. Some jurisdictions prescribe or allow in minor cases a jury of six. Oregon allows 10:2 verdicts—that is, a majority of 10—in all criminal cases, except capital ones; and in 1968 England followed the Oregon example. A few Southern states in the United States allow majority verdicts in misdemeanour trials. In civil cases many states now allow 10:2 verdicts. When the required number (12 or 10) of jurors cannot agree on a verdict (termed a hung jury in the United States), the judge declares a mistrial, which means the case, unless it is withdrawn, must be tried anew. It is somewhat remarkable that "hung" juries occur with relative infrequency even when unanimity is required. In Europe juries operate under a different principle. Unless at least two-thirds of all the jurors vote guilty, the defendant must be acquitted. The U.S. Army court-martial jury also operates under this principle.

Sentencing. Although in civil cases the jury decides both issues of liability and amount of damages, in criminal cases it has been restricted generally to the issues of guilt, while punishment has been left to the judge. In some Southern U.S. states, however, the jury also decides the sentence within a certain range that the law provides. And, in all jurisdictions that have retained the death penalty, if the jury finds the defendant guilty of the capital crime, it decides, or at least expresses an opinion, as to whether the death penalty is to be imposed. In most jurisdictions decisions on guilt and sentence are rendered simultaneously, but California has introduced the so-called second trial in capital cases, which occurs after a guilty verdict. At such a "second trial" pleas and evidence are presented for and against the imposition of the death penalty in the specific case, and only then is the jury asked to determine the sentence.

Relation of judge and jury

Control. Trial by jury is, of course, trial by jury under the supervision of a judge. The formula for sharing power between judge and jury is complex. First, the judge decides what the jury may or may not hear under the rules of evidence. Second, if the judge finds that the evidence presented leaves no factual issue to be resolved, he may withdraw the issue from the jury and direct the jury to acquit a defendant or, in a civil trial, find for either plaintiff or defendant; he cannot, however, direct a guilty verdict in a criminal trial. Third, in some jurisdictions the judge may, and often will, summarize the evidence or even discuss its weight. Fourth, the judge instructs the jury as to the law it should apply in reaching the verdict. Finally, if the judge finds the jury's verdict to be manifestly against the weight of the evidence, he may with one exception set it aside and order a new trial. The only exception is in a criminal case in which the jury renders an acquittal; under Anglo-American law (though not under European continental law) the jury's acquittal is always final.

The jury normally renders a general verdict—that is, a yes or no answer to liability or guilt—and does not give reasons for its decision. At times, however, courts employ "special verdicts" or "special interrogatories" in which the jurors are asked to decide a series of specific factual issues that bear on the overall verdict.

The controversy over the jury. The jury has been enmeshed in a perennial debate as to its merits, a debate that has recruited some of the great names in law and political philosophy—from Montesquieu, William Blackstone, and Thomas Jefferson to present-day theorists and practitioners—and has centred on three issues. First, there is the debate about collateral aspects: there are favourable contentions that the jury provides an important civic experience, that it makes tolerable the stringency of certain decisions, that it acts as a sort of lightning rod for animosity that otherwise might centre on the more permanent judge, and that the jury is a guarantor of integrity since it is said to be more difficult to bribe 12 people than one. Against this it has been urged that jury duty disenchants the citizen, that it imposes an unfair burden, that the jury is expensive, and that it makes it difficult to do away with the often interminable delays that exist in civil litigation.

The jury's competence

Second, there is the issue of the jury's competence. It is argued that the judge, by training, discipline, experience, and superior intelligence, is better able to understand law and facts than laypersons drawn from a broad range of levels of intelligence, without experience and without durable official responsibility. But it is also argued that 12 heads are better than one, that the jury as a group has wisdom and strength beyond that of its individual members, that it makes up in common sense and experience what it lacks in training, and that its very inexperience is an asset because it secures a fresh perception of each trial, avoiding the stereotypes that may infect the judicial eye.

Finally, there is the question of the jury's interpretation of the law. The critics complain that the jury will not follow the law, either because it does not understand it or because it does not like it, and hence will administer justice unevenly and that the jury produces a government by individuals and not by rule of law, against which Anglo-American political tradition is so steadfastly set. The jury's champions offer this very flexibility as its most endearing characteristic. They see the jury as a remarkable device for ensuring that the rigidity of the general rule can be shaped to justice in a particular case, with government by the spirit of the law and not by its letter.

Performance. In a recent survey of some 7,000 jury trials, the presiding judges were requested to reveal how they would have decided without a jury; the results of the survey provided some major insights into the actual performance of the contemporary American jury. In both civil and criminal trials, judge and jury agreed in 78 percent of all verdicts. In civil cases the disagreement in the remaining cases was symmetrically split; in 19 percent of the criminal cases, however, the judge would have convicted, whereas the jury acquitted. The letter of the law confines the jury to "finding the facts," but the deviations from the judge are mostly due to the jury's subtle, and not always conscious, injecting its sense of justice into a case that might go either way. This sense of justice may be concerned with the person of the accused, with the threat of too harsh a punishment, or with the content of the criminal law rules. Thus, close study of the jury has revealed it as a highly sensitive institution, subtle and discerning, moved by factors far beyond gross sympathy for the defendant. On the whole, the system tolerates and even appreciates these deviations of the jury from the judge, even if in rare cases they reflect what the national community experiences as intolerable local prejudice.

(H.Ka./H.Z./Ed.)

OTHER JUDICIAL OFFICIALS

In most countries there are other officials who serve the court. Court clerks, who are responsible for case records and documents, and bailiffs, who are in charge of keeping order, are found in most judicial systems. Also prevalent are officers who prosecute cases in the government's name: states attorneys and district attorneys in the United States, procurators-general in the U.S.S.R., and *procureurs généraux* in France.

Probation officers are found in many countries including the U.S. and Japan. Notaries in France, Italy, and the U.S.S.R. have greater powers than their counterparts in the U.S. In fact, they perform many services carried out by lawyers in the common-law system, such as drafting and verifying wills and contracts and preparing petitions for presentation in court.

Certain countries have officials that are particularly indigenous to their country or legal system. France, for example, has a *juge d'instruction,* who is responsible for the preliminary investigative proceedings prior to a criminal trial.

THE STRUCTURE AND STATUS OF THE JUDICIARY UNDER COMMUNISM

Although the essential legal institutions of the Soviet Union and other Communist countries are based on the civil-law system, certain features are unique. These characteristics are partly the result of the Soviet Union's attitudes toward law that antedate the Soviet system, but mostly they result from the attempt to reconcile Marxist theory with the institutional needs of a modern society.

According to Marx and his followers, the legal system, like all other governmental structures and instruments of class oppression, would "wither away" in a Communist society; thus the courts that existed after the Revolution were considered temporary institutions, required only during the transition to Communism. The ordinary and traditional business of the courts was carried on by the so-called people's courts, while "revolutionary tribunals" dealt with individuals the government considered to be political opponents. A nonjudicial body in the hands of the secret police (at first called Cheka, later OGPU and NKVD), operating in the style of an administrative agency, also heard cases and handed out sentences—usually of the severest kind.

The Soviet three-level court system

In 1921 some capitalist measures were temporarily introduced to revive the economy, and this necessitated some stabilization of the legal system and its institutions. A three-level system of courts with civil and criminal jurisdiction was established in 1922 for the Russian Republic, which in the same year formed a federation with the other soviet republics under its jurisdiction, making up the Union of Soviet Socialist Republics. A new constitution created a federal court—the U.S.S.R. Supreme Court—and a federal judiciary act of 1924 established uniform principles for the judiciary throughout the republics, patterned largely after the system adopted by the Russian Republic. The basic structure of the courts laid down at that time has remained essentially the same to the present, with some minor changes and reforms.

The "People's Courts" on the local level are courts of original jurisdiction for minor criminal cases and a large number of civil cases. The next level, the provincial courts, receive appeals from the people's courts and have original jurisdiction over political and serious civil and criminal cases. The highest level in each republic is its supreme court, which hears appeals from the provincial courts, disciplines lower courts, and has some original jurisdiction over extremely serious cases.

On all three levels, appellate cases are tried by a court consisting of three full-time judges, whereas one judge and two lay judges, or assessors, preside over cases on first hearing. Judges of the people's courts are popularly elected every five years, and judges on the provincial and supreme court levels are "elected" by soviets (bodies combining legislative and executive functions) of the corresponding levels of government. All judges may be recalled before the expiration of their terms by those who elected them.

The federal court system is twofold. There are courts called military tribunals that deal with charges against men in the armed forces and with charges of espionage brought against civilians. The other federal body is the U.S.S.R. Supreme Court—the highest judicial body—which has original jurisdiction in a few special cases relating to the survival of the regime, appellate power over the decisions of the supreme courts of the republics or decisions of the military tribunals, and the right to issue directives to all inferior courts in matters of administration of justice on the basis of a series of its decisions. Although Soviet legal theory is patterned after that of civil-law countries in that it does not recognize judicial lawmaking, these Supreme Court directives function as a source of law and are binding on all courts.

The status of judges

The status of the judiciary in the Soviet Union has undergone some changes that parallel the institutional changes since the early days of the Revolution. The system organized in 1922 had the stated purpose of safeguarding the conquests of the Revolution and establishing the dictatorship of the proletariat. Judges were called upon to use their "revolutionary conscience" in deciding cases, and the doctrine of impartiality and independence of the judiciary was repudiated. With the passage of time, however, the Soviet rulers found the need for legal institutions of a stable nature increasing rather than decreasing, and the goal of the legal system was changed from protection of a particular class to protection of the socialist order and the rights of all citizens. Although the role of the judiciary is still conceived of as a political task, there is some acceptance of the idea that judges should be independent and impartial. Marxist philosophy notwithstanding, the Soviet Union and other Socialist countries are confronted with a growing need for legal institutions to fulfill many of the same functions as those in the West. One attempt to fill this need in recent years has been the appearance of "social organizations," such as the "comrades' courts," which are described as voluntary organizations using persuasion and social influence to deal with matters that would otherwise come before a court. But these organizations are party controlled, have only limited power to impose sanctions, and do not appear at present to offer an effective alternative to the type of legal institutions that have been developing within Soviet society.

The other Communist countries, both in eastern Europe and Asia, adopted legal institutions patterned largely after the Soviet model. Since Stalin's death, however, there have been some modifications in the eastern European countries, coinciding with the reforms in the civil and criminal codes adopted by the Soviet Union in the late 1950s and early 1960s. Chinese leaders, however, have resisted efforts to codify their laws, preferring flexibility in their courts, and they have abandoned the policy of copying Soviet legal patterns.

(Ed.)

Arbitration

Arbitration is a nonjudicial, legal technique for resolving disputes by referring them to a third party for a binding decision, or "award," as an arbitrator's findings are usually described. The arbitrator may be a single person or an arbitration board, usually of three members. Arbitration is most commonly resorted to for the resolution of commercial disputes and must be distinguished from mediation and conciliation, which are common in the settlement of labour disputes between management and labour unions. In such cases, the parties resort to a third person to offer a recommendation for a settlement or to help them to reach a compromise. Such intervention by a third party, which also occurs in international disputes between states in the form of diplomatic intervention and good offices, has no binding force upon the disputants, as has the arbitrator's decision, the award.

COMMERCIAL ARBITRATION

Commercial arbitration is a means of settling disputes by referring them to a third person, an arbitrator, selected by the parties for a decision based on the evidence and arguments presented to the arbitration tribunal. The parties agree in advance that the decision will be accepted as final and binding upon them.

Historically, commercial arbitration was used in resolving controversies between medieval merchants, in fairs and marketplaces in England and on the European continent, and in the Mediterranean and Baltic sea trade. The increased use of commercial arbitration became possible after courts were empowered to enforce the parties' agreement to arbitrate. The first such statute was the English Arbitration Act of 1889, now consolidated into an act of 1950 and adopted by arbitration statutes in most countries of the Commonwealth. It was followed in the United States by an arbitration statute of the state of New York in 1920 and a Federal Arbitration Act of 1925. Codified in 1940, the latter deals with the enforcement in federal courts of arbitration agreements and awards in maritime transactions and those involving interstate and foreign commerce. Most states of the United States adopted, sometimes with minor changes, the Uniform Arbitration Act of 1955, as amended in 1956, which had been promoted by the Commissioners on Uniform State Laws and recommended by the American Bar Association. This act provides for the judicial enforcement of an agreement to arbitrate existing and future disputes, thereby making the arbitration agreement no longer revocable, as it had been under common law. It also provides for the substitution of arbitrators in the event of a party failing to select an arbitrator and for a suspension of any court action instituted in contravention of a voluntary arbitration agreement. The courts thereby play an important role in implementing arbitration agreements, making judicial assistance available against a recalcitrant party. This concept of modern arbitration law, which recognizes the irrevocability of arbitration agreements and the enforceability of awards prevails also in the arbitration statutes of nearly all countries of Europe and Asia. Latin American procedural laws generally provide only for court enforcement of agreements to arbitrate existing disputes and do not provide for the enforcement of subsequent disputes that may arise under the arbitration agreement.

Statutory development of arbitration

Function and scope. Arbitration has been used customarily for the settlement of disputes between members of trade associations and between different exchanges in the securities and commodities trade. Form contracts contain a standard arbitration clause referring to the arbitration rules of the respective organization. Numerous arrangements between parties in industry and commerce also provide for arbitration of controversies arising out of contracts for the sale of manufactured goods, for terms of service of employment, for construction and engineering projects, for financial operations, for agency and distribution arrangements, and for many other undertakings.

The usefulness and significance of arbitration is demonstrated by its increasing use by the business community and the legal profession in many countries of the world. The primary advantage is the speed with which controversies can be resolved by arbitration, compared with the long delays of ordinary court procedure. The expert knowledge of arbitrators of the customs and usages of a specific trade makes testimony by others and much documentation unnecessary, thereby eliminating expenses connected with court procedures. The privacy of the arbitration procedure is also much valued by parties to the controversy; situations unfavourable to the party's credit or deficiencies in manufactured goods revealed in arbitration proceedings do not become known to outsiders. There are, however, also disadvantages in the arbitration process. The fact that in Anglo-American practice no reasons are given by the arbitrator to accompany his award prevents the development of a guideline for the further conduct of business relations. This uncertainty resulting from lack of reasoned precedents, moreover, makes the arbitral decision less predictable. Further obstacles to the wider use of commercial arbitration are the divergencies in municipal laws and court decisions that result in different interpretations of similar arbitration questions and the fact that awards are not published: publication of awards, even without identification of the parties, might assist in the establishment of precedents useful in discouraging future disputes on similar issues in a specific branch of industry or commerce.

Advantages of arbitration over court procedure

Procedure. The method of selecting arbitrators is an important aspect of the arbitration process, for the arbitrator's ability and fairness is the decisive element in any arbitration. The general practice is for both parties to select an arbitrator at the time a conflict arises or at the time the arbitration agreement is concluded. The two arbitrators then select a chairman, forming a tribunal. Selection of arbitrators is also often made by agencies administering commercial arbitration under preestablished rules of procedure. These organizations—various trade associations, produce exchanges, and chambers of commerce in many countries—maintain panels of expert arbitrators. The parties may either make their own selection or entrust the appointment of the arbitrators to the organization.

Challenges to the arbitration process are not uncommon. A party may claim, for example, that no valid arbitration agreement came into existence because the person signing the agreement had no authority to do so or that a condition precedent to arbitration had not been fulfilled. More often, arbitration is contested on the ground that the specific controversy is not covered by the agreement. In such cases, the issue of whether the arbitrator has authority to deal with the conflict is usually determined by a court. Further challenges to the arbitration process may be directed against an arbitrator, on grounds, for example, of alleged lack of impartiality. Any such challenge can generally be maintained only after the arbitration has been concluded, as courts are reluctant to interfere with the arbitration process before an award has been rendered.

The arbitration process is governed by the rules to which the parties referred in their agreement; otherwise, the procedure will be determined by the arbitrators. The arbitration proceedings must be conducted so as to afford the parties a fair hearing on the basis of equality. The arbitrator generally has the authority to request the parties and third persons to produce documents and books and to enforce such a request by issuing subpoenas. If a party fails to appear at a properly convened hearing, without showing a legitimate cause, the arbitrator in most instances will proceed in the absence of that party and render an award after investigation of the matter in dispute.

Under the law and arbitration practice of most countries, the award is valid and binding upon the parties when rendered by a majority of the arbitrators, unless the parties expressly request a unanimous decision of the arbitrators, which they seldom do. The statutory law of various countries and the rules of agencies administering commercial arbitration contain provisions on the form, certification, notification, and delivery of the award, with which requirements the arbitrator has to comply.

A much-disputed question in commercial arbitration concerns the law to be applied by the arbitrators. Generally, the award must be based upon the law as determined by the parties in their agreement. This failing, the arbitrator must apply the law he considers proper in accordance with the rules of conflict of laws. In both cases, the arbitrator will have to take account of the terms of the contract and the usage of the specific trade. If, during any arbitration proceeding, a compromise is reached by the parties, that compromise may be recorded as an award by the arbitrator.

Appeal of arbitration decisions to the courts

Appeals to the courts from the award cannot be excluded by agreement of the parties, since the fairness of the arbitration process as a quasi-judicial procedure has to be maintained. Any court control is, however, confined to specific matters, usually enumerated in the arbitration statutes, such as misconduct of the arbitrator in denying a party the full presentation of its claim or not granting a postponement of the hearing for good cause. A review of the award by a court generally will not deal with the arbitrator's decisions as to facts or with his application of the law. The competence of the courts is restricted in order not to make the arbitration process the beginning of litigation instead of its end. Recognition of an award and its enforcement will be denied when it appears to be contrary to public policy, as might be the case, for example, in cases involving trusts (monopolies), industrial property rights, or violation of foreign-currency restrictions. An arbitration award has the authority of a court decision and may be enforced by summary court action according to the procedural law of the country in which execution is being sought.

International commercial arbitration. International commercial arbitration between traders of different countries has long been recognized by the business community and the legal profession as a suitable means of settling trade controversies out of court. The procedure in international commercial arbitration is basically the same as in domestic arbitration. In order to establish more uniformity in procedure and to make access to arbitration facilities more easily available, the United Nations economic commissions in 1966 published new rules applying to international arbitration. Those for Europe are contained in the "Arbitration Rules of the United Nations Economic Commission for Europe" and for Asia and the Far East in the "Economic Commission for Asia and the Far East Rules for International Commercial Arbitration."

The development of international commercial arbitration has been furthered by uniform arbitration legislation prepared by the United Nations Conference on International Commercial Arbitration in 1958 and by the Council of Europe and the Inter-American Juridical Committee of the Organization of American States. One particularly difficult problem of international commercial arbitration is the enforcement of awards in a country other than the one in which they were rendered. Statutory municipal laws do not usually contain provisions for the enforcement of foreign awards, and parties are faced with uncertainty about the law and practice of enforcement procedure in a country other than their own.

The aforementioned international agreements, to which most of the trading nations of the world adhere, facilitate the enforcement of foreign awards to the extent that no further action is necessary in the country in which the award was rendered: the opposing debtor must establish that the award has been set aside or that its effects have been suspended by a competent authority, thus shifting the burden of proof of the nonbinding character of the award to the losing party.

Further development of international commercial arbitration is encouraged by the United Nations Commission on International Trade Law, which aims at promoting the harmonization and unification of laws in the field of international commercial arbitration.

LABOUR ARBITRATION

Labour arbitration—the reference of disputes between management and labour unions to an impartial third party for a final resolution—is usually the last step under a collective-bargaining agreement after all other measures to achieve a settlement have been exhausted. Labour arbitration is not, as is commercial arbitration, an auxiliary avenue of justice and thereby a substitute for ordinary court procedure but a technique also for settling or avoiding strikes.

Two major aspects of labour arbitration

Two major aspects of labour arbitration are usually distinguished: arbitration of rights and arbitration of interests. Arbitration of rights refers to the arbitration of an existing labour contract when a dispute over the application of that contract arises between labour and management. Arbitration of interests refers to arbitration between labour and management during the negotiation of a new labour contract.

Arbitration of rights. Arbitration of rights under the terms of a collective-bargaining agreement is employed in the United States far more than in most other countries. Outside the United States, labour courts, industrial courts, or conciliation and arbitration commissions perform the function of arbitrating rights. These bodies are usually appointed by the government, and recourse to them is frequently compulsory.

More than 90 percent of the collective-bargaining agreements in the United States provide for arbitration as a last step in the grievance procedure. Employees, for example, through their union, may present for arbitration complaints concerning such matters as discipline, discharge, and violations of working conditions. Other issues frequently submitted to arbitration customarily concern premium payments and incentive rates, overtime and vacations, Christmas bonuses, seniority rights, and fringe benefits, such as pension and welfare plans.

The arbitrator's decision must be based on the collective-bargaining agreement, which provides for the application of an existing contract to the grievance presented. The question of whether the various steps in the grievance procedure have been complied with before the initiation of the arbitration is usually left to the determination of the arbitrator and not of a court. The question, however, of whether the disputed issue is covered by the collective-bargaining agreement has to be determined by a court and not by the arbitrator. This authority of the courts was upheld by the Supreme Court of the United States in 1960.

Selection of labour arbitrators

The choice of arbitrator is made either by naming him in the agreement or, more often, by leaving the choice open until a dispute has arisen. Frequently, only a single arbitrator is appointed—usually an expert in the field of industrial relations. Alternatively, tripartite arbitration boards are established, both parties appointing their own arbitrator, who acts somewhat as advocate. A neutral chairman is selected either by the parties or by the two party-appointed arbitrators.

A further technique of arbitration of rights is the appointment of a single permanent arbitrator, or "umpire," to resolve disputes for the duration of the collective-bargaining agreement. The umpire will be intimately acquainted with the various economic, financial, and other aspects of the particular industry and will be familiar with the relationship between management and union developed in the past. He sometimes follows precedents, especially those established by his predecessor. This permanent umpire system originated in the United States in the anthracite-coal industry at the beginning of the 20th century and has been employed in such other important industries as newspaper printing and clothing.

Labour arbitrators render binding decisions on the disputes submitted to them. They are not bound by strict rules of court procedure, especially as regards burden of proof and the presentation of evidence. As arbitrators, they have the power to subpoena persons and written evidence. Labour arbitrators tend to evaluate factual evidence rather freely and often reduce penalties imposed upon employees by the management for breach of the labour contract. Even minor questions, such as the use of company time by employees for washups or coffee breaks, are submitted to arbitration, in order to establish precedents in the operation of the plant. Generally, however, arbitrators are not bound to follow previous decisions.

Decisions of labour arbitrators are seldom reviewed by the courts, as awards are usually fully complied with by both parties.

Arbitration of interests. Arbitration of the terms of a new contract, referred to as arbitration of interests, may be instituted if management and the labour union are unable to agree on a new contract. In some industries, such as hotels, printing, transit, and utilities, such disputes are submitted to arbitration. In most countries, however, management and union are seldom inclined to forgo resort to lockouts and strikes in an attempt to obtain favourable new contracts, and interest arbitration is thus seldom used.

Compulsory arbitration, directed by legislative fiat, has been a controversial issue in the settlement of industrial disputes. It has been favoured in disputes in the transportation industry, which may involve great public inconvenience, and in disputes in the public-utilities sector when an immediate danger to public health and safety might occur. Compulsory arbitration has been declared unconstitutional in some states of the United States. More recently, however, it has been adopted as a regular procedure for the settlement of disputes with municipal employees in some U.S. cities.

INTERNATIONAL ARBITRATION

Controversies between sovereign states that are not settled by diplomatic negotiation or conciliation are often referred, by agreement of both parties, to the decision of a third disinterested party, who arbitrates the dispute with binding force upon the disputant parties. Such arbitration between states has a long history: it was used between city-states in ancient Greece and also in the Middle Ages, when the pope often acted as the sole arbitrator.

Historical development. The modern development of international arbitration can be traced to the Jay Treaty of 1794 between Great Britain and the United States, which established three arbitral commissions to settle questions and claims arising out of the American Revolution. In the 19th century many arbitral agreements were concluded by which arbitration tribunals were established ad hoc to deal with a specific case or to handle a great number of claims. Most significant was the "Alabama" claim arbitration under the Treaty of Washington (1871), by which the United States and Great Britain agreed to settle claims arising from the failure of Great Britain to maintain its neutrality during the U.S. Civil War.

Commissions consisting of members drawn from both disputant countries ("mixed arbitral commissions") were often used in the 19th century to settle pecuniary claims for compensation of injuries to aliens for which justice could not be obtained in foreign courts. Such was the purpose of a convention of 1868 between the United States and Mexico, by which claims of citizens of each country arising from the Civil War were settled. Boundary disputes between states were also often settled by arbitration.

Development of international arbitration procedures

International arbitration was given a more permanent basis by the Hague Conference of 1899, which adopted a convention on the pacific settlement of international disputes, revised by a Conference of 1907. The convention stated that:

> International arbitration has for its object the settlement of disputes between States by judges of their own choice and on the basis of respect for law. Recourse to arbitration implies an engagement to submit in good faith to the award.

A Permanent Court of Arbitration, composed of a panel of jurists appointed by the member governments, from which the litigant governments may select the arbitrators, was established at The Hague in 1899.

Twenty cases were arbitrated between 1902 and 1932, but from that year until 1972 only five cases were dealt with. This was largely because the importance of the Permanent Court of Arbitration was lessened by the Permanent Court of Justice (established in 1922) and its successor, the International Court of Justice. More recently, in 1960, the court, which was originally devised for the settlement of disputes between states, has offered its services for the arbitration of controversies between states and individuals or corporations. Such a dispute was arbitrated in 1970 between a British company and the government of the Democratic Republic of Sudan. The case concerned the repudiation of a contract for building houses in the irrigation zone of the Khashm al-Qirbah Dam in The Sudan.

Arbitration provisions of international treaties. There are several multilateral treaties that provide for the pacific settlement of international disputes by arbitration, including the Geneva General Act for the Settlement of Disputes of 1928, adopted by the League of Nations and reactivated by the General Assembly of the United Nations in 1949, which provides for the settlement of various disputes, after unsuccessful efforts at conciliation, by an arbitral tribunal of five members. Other such treaties include the General Treaty of Inter-American Arbitration, signed in Washington in 1929, and the American Treaty on Pacific Settlement of Disputes, signed in Bogotá in 1948. More recently the Council of Europe adopted the European Convention for the Peaceful Settlement of Disputes (1957).

Arbitration is also mentioned as a proper method of settling disputes between countries in the Charter of the United Nations, as it was in the Covenant of the League of Nations.

Model rules of the International Law Commission

The International Law Commission of the United Nations submitted to the General Assembly in 1955 a Convention on Arbitral Procedure. Its model rules would not become binding on any member-state of the United Nations unless they were accepted by a state in an arbitration treaty or in a special arbitral agreement. The model rules, however, have not yet been adopted in any arbitration arrangement between disputant governments, although in 1958 the General Assembly recommended the model rules for use by member-states when appropriate. It seems clear that states prefer flexibility in the resolution of their disputes by arranging the rules and proceedings of an arbitration according to circumstances.

Great impediments, indeed, still exist in the acceptance of international arbitration, especially in cases in which disputes between governments and foreign private parties are involved. In such cases the state will often insist that its own local remedies—administrative and court proceedings—be exhausted. Generally, the government of the national who advances a claim against a foreign government will require evidence that the injured party has pursued all remedies in the foreign country before it presses a claim for international negotiation and adjudication, if, indeed, it decides to take up the case at all. Contracting parties may agree in their contract that they need not exhaust local remedies before resorting to arbitration, and one 1965 instrument, the Convention on the Settlement of Investment Disputes, states:

> Consent of the parties to arbitration under this Convention shall, unless otherwise stated, be deemed consent to such arbitration to the exclusion of any other remedy. A Contracting State may require the exhaustion of local administrative or judicial remedies as a condition of its consent to arbitration under this Convention.

The arbitration agreement in a general multilateral treaty, a bilateral convention, or in a specific contractual arrangement between two states often does not deal with the selection of the arbitrators and the appointment of an umpire, the procedure to be followed in the arbitration, the subject matter of the dispute, the specific issues to be submitted, the presentation of evidence, the place of the hearings, the law to be applied by the arbitrators, and the time when the award has to be rendered. These questions are usually dealt with in an agreement between the parties to the dispute known as compromis. If the compromis fails in some particular—to define the applicable law, for example—it is generally understood that the arbitrator shall apply international law.

An award rendered by an arbitral tribunal is customarily complied with by states: it is, in fact, invariably the case that unless a state is prepared to comply with an adverse decision, it will not submit the dispute to arbitration. The difficulties in the use of international arbitration thus consist less in the enforcement of arbitral awards than in persuading states involved in disputes to submit them to a third party, an arbitrator, or an arbitration tribunal.

(M.Do.)

BIBLIOGRAPHY. Overviews of legal institutions within specific countries include R.M. JACKSON, *The Machinery of Justice in England*, 7th ed. (1977); LEWIS MAYERS, *The American Legal System*, rev. ed. (1964, reprinted 1981); ROBERT A. CARP and RONALD STIDHAM, *Judicial Process in America*, 2nd ed. (1993),

from a political perspective; G. LEROY CERTOMA, *The Italian Legal System* (1985); and HARIHAR PRASAD DUBEY, *A Short History of the Judicial Systems of India and Some Foreign Countries* (1968). Works with a broader scope are RENÉ DAVID and JOHN E.C. BRIERLEY, *Major Legal Systems in the World Today*, 3rd ed. (1985; originally published in French, 8th ed., 1982); SYBILLE BEDFORD, *The Faces of Justice* (1961), discussing how cases are handled in England, Germany, Austria, Switzerland, and France; JOHN HENRY MERRYMAN, *The Civil Law Tradition: An Introduction to the Legal Systems of Western Europe and Latin America*, 2nd ed. (1985), summarizing the principles and institutions in civil-law countries; and C.G. WEERAMANTRY, *Islamic Jurisprudence* (1988), concentrating on Islāmic issues. KARL LLEWELLYN, *The Case Law System in America* (1989; originally published in German, 1933), is a pre-World War II classic on the American legal system from a foreign perspective.

Both JAMES E. BOND, *The Art of Judging* (1987); and STEVEN J. BURTON, *Judging in Good Faith* (1992), provide a philosophical approach to law judging. JOEL LEVIN, *How Judges Reason: The Logic of Adjudication* (1992), explains the author's theory of judicial pluralism and criticizes other theories. LAWRENCE M. SOLAN, *The Language of Judges* (1993), offers a linguistic analysis of judicial language. HENRY J. ABRAHAM, *The Judicial Process: An Introductory Analysis of the Courts of the United States, England, and France*, 6th ed. (1993), is a classic text. BENJAMIN N. CARDOZO, *The Nature of the Judicial Process* (1921, reissued 1991), is an explanation by a distinguished judge of how an appellate court reaches its decisions. ROSCOE POUND, *Organization of Courts* (1940, reprinted 1979), gives a detailed treatment of U.S. court structure. W.R. CORNISH, *The Jury*, rev. ed. (1971), is a comprehensive British essay on the jury, combining traditional learning with new empirical material. OLIVER WENDELL HOLMES, JR., *The Common Law* (1881, reprinted 1991), is a classic treatment of the growth of law through judicial decisions.

Discussions of politics and law include JEROME R. CORSI, *Judicial Politics* (1984), an introductory text; DAVID KAIRYS (ed.), *The Politics of Law: A Progressive Critique*, rev. ed. (1990), a compilation of essays; and JEROLD L. WALTMAN and KENNETH M. HOLLAND (eds.), *The Political Role of Law Courts in Modern Democracies* (1988), examining the United States, Australia, Canada, the former West Germany, England, France, Italy, Sweden, and Japan. PATRICIA SMITH (ed.), *Feminist Jurisprudence* (1993), is a compilation of essays by leading scholars. HARRY H. WELLINGTON, *Interpreting the Constitution: The Supreme Court and the Process of Adjudication* (1990), is a liberal review of the Supreme Court that rejects original meaning by presenting the Constitution as a living document.

The leading treatises and handbooks on commercial arbitration in English are *Russel on the Law of Arbitration*, 20th ed. by ANTHONY WALTON and MARY VITORIA (1982); MARTIN DOMKE, *The Law and Practice of Commercial Arbitration* (1968); NRISINHADAS BASU, *The Arbitration Act, 1940 (X of 1940)*, ed. by SUDHIR KUMAR BOSE, 5th ed. (1965); and the INTERNATIONAL ASSOCIATION OF LAWYERS, *International Commercial Arbitration*, ed. by PIETER SANDERS, 3 vol. (1956–65).

Labour arbitration is treated in FRANK ELKOURI and EDNA ASPER ELKOURI, *How Arbitration Works*, 4th ed. (1985); and CLARENCE M. UPDEGRAFF and WHITLEY P. MCCOY, *Arbitration of Labor Disputes*, 2nd ed. (1961). Also helpful is EDWIN EMIL WITTE, *Historical Survey of Labor Arbitration* (1952), describing the historical development of arbitration. Detailed primers are ARNOLD M. ZACK, *A Handbook for Grievance Arbitration: Procedural and Ethical Issues* (1992); and CHARLES S. LACUGNA, *An Introduction to Labor Arbitration* (1988). TIA SCHNEIDER DENENBERG and R.V. DENENBERG, *Alcohol and Other Drugs: Issues in Arbitration* (1991), focuses on this growing problem. MAX ZIMNY, WILLIAM F. DOLSON, and CHRISTOPHER A. BARRECA (eds.), *Labor Arbitration: A Practical Guide for Advocates* (1990), is both a training guide and a resource volume. DAVID B. LIPSKY and CLIFFORD B. DONN (eds.), *Collective Bargaining in American Industry* (1987), provides examples from specific industries. ARNOLD M. ZACK (ed.), *Arbitration in Practice* (1984), examines specific aspects of the arbitration process. WALTER E. BAER, *Arbitration for the Practitioner* (1988), is a guide for those involved in arbitration. DOUGLAS M. MCCABE, *Corporate Nonunion Complaint Procedures and Systems: A Strategic Human Resources Management Analysis* (1988), examines arbitration and grievance procedures in nonunion companies.

Various aspects of international arbitration are dealt with by J.L. SIMPSON and HAZEL FOX, *International Arbitration: Law and Practice* (1959); DAVID DAVIES MEMORIAL INSTITUTE OF INTERNATIONAL STUDIES, *Report of a Study Group on the Peaceful Settlement of International Disputes* (1966); PIETER SANDERS (ed.), *International Arbitration* (1967); and MAURO RUBINO-SAMMARTANO, *International Arbitration Law* (1990). T. HANAMI and B. BLANPAIN (eds.), *Industrial Conflict Resolution in Market Economies: A Study of Canada, Great Britain, and Sweden* (1987), and *Industrial Conflict Resolution in Market Economies: A Study of Australia, the Federal Republic of Germany, Italy, Japan, and the USA*, 2nd ed. (1989), provide general overviews and examples for each country from both public and private sectors. CHRISTINE D. GRAY, *Judicial Remedies in International Law* (1987), includes treatment of international arbitration tribunals. ALAN REDFERN, MARTIN HUNTER, and MURRAY SMITH, *Law and Practice of International Commercial Arbitration*, 2nd ed. (1991), is an extensive treatise.

Works on other types of law include MELVIN ARON EISENBERG, *The Nature of the Common Law* (1988), discussing common-law decisionmaking; P.S.R.F. MATHIJSEN, *A Guide to European Community Law*, 5th ed. (1990), a general introduction; and DELMAR KARLEN, GEOFFREY SAWER, and EDWARD M. WISE, *Anglo-American Criminal Justice* (1967), comparing two judicial systems in the area of criminal law.

(D.K./H.Ka./H.Z./M.Do./Ed.)

Kant and Kantianism

Immanuel Kant was the foremost thinker of the Enlightenment and one of the great philosophers of all time, in whom were subsumed new trends that had begun with the Rationalism (stressing reason) of René Descartes and the Empiricism (stressing experience) of Francis Bacon. He inaugurated a new era in the development of philosophical thought. His comprehensive and systematic work in theory of knowledge, ethics, and aesthetics greatly influenced all subsequent philosophy, especially the various German schools of Kantianism and Idealism.

This article deals with the man, his achievements, and the subsequent history of Kantianism. It is divided into the following sections:

Life and works

BACKGROUND AND EARLY YEARS

Kant was born on April 22, 1724, at Königsberg in East Prussia (now Kaliningrad, Russia) and lived in that remote province for his entire life. His father, a saddler, was, according to Kant, a descendant of a Scottish immigrant, although scholars have found no basis for this claim; his mother, an uneducated German woman, was remarkable for her character and natural intelligence. Both parents were devoted followers of the Pietist branch of the Lutheran Church, which taught that religion belongs to the inner life expressed in simplicity and obedience to moral law. The influence of their pastor made it possible for Kant—the fourth of nine children, but the eldest surviving child—to obtain an education.

Pietist rearing and schooling

At the age of eight Kant entered the Pietist school that his pastor directed. This was a Latin school, and it was presumably during the eight and a half years he was there that Kant acquired his lifelong love for the Latin classics, especially for the naturalistic poet Lucretius. In 1740 he enrolled in the University of Königsberg as a theological student. But, although he attended courses in theology and even preached on a few occasions, he was principally attracted to mathematics and physics. Aided by a young professor who had studied Christian Wolff, a systematizer of Rationalist philosophy, and who was also an enthusiast for the science of Sir Isaac Newton, Kant began reading the work of the English physicist and, in 1744, started his first book, dealing with a problem concerning kinetic forces. Though by that time he had decided to pursue an academic career, the death of his father in 1746 and his failure to obtain the post of undertutor in one of the schools attached to the university compelled him to withdraw and seek a means of supporting himself.

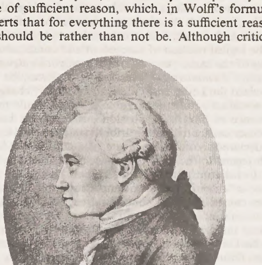
Tutor and Privatdozent. He found employment as a family tutor and, during the nine years that he gave to it, worked for three different families. With them he was introduced to the influential society of the city, acquired social grace, and made his farthest travels from his native city—some 60 miles (96 kilometres) away to the town of Arnsdorf. In 1755, aided by the kindness of a friend, he was able to complete his degree at the university and take up the position of *Privatdozent,* or lecturer.

The three early dissertations

Three dissertations that he presented on obtaining this post indicate the interest and direction of his thought at this time. In one, *De Igne* (On Fire), he argued that bodies operate on one another through the medium of a uniformly diffused elastic and subtle matter that is the underlying substance of both heat and light. His first teaching was in mathematics and physics, and he was never to lose his interest in scientific developments. That it was more than an amateur interest is shown by his publication within the next few years of several scientific works dealing with the different races of men, the nature of winds, the causes of earthquakes, and the general theory of the heavens.

At this period Newtonian physics was important to Kant as much for its philosophical implications as for its scientific content. A second dissertation, the *Monodologia physica* (1756), contrasted the Newtonian methods of thinking with those employed in the philosophy then prevailing in German universities. This was the philosophy of Gottfried Wilhelm Leibniz, a universal scholar, as systematized and popularized by Wolff and by Alexander Gottlieb Baumgarten, author of a widely used text, the *Metaphysica* (1739). Leibniz' works as they are now known were not fully available to these writers; and the Leibnizian philosophy that they presented was extravagantly Rationalistic, abstract, and cut-and-dried. It nevertheless remained a powerful force, and the main efforts of independent thinkers in Germany at the time were devoted to examining Leibniz's ideas.

In a third dissertation, *Principiorum Primorum Cognitionis Metaphysicae Nova Dilucidato* (1755), on the first principles of metaphysics, Kant analyzed especially the principle of sufficient reason, which, in Wolff's formulation, asserts that for everything there is a sufficient reason why it should be rather than not be. Although critical,



Kant, pencil portrait by Hans Veit Schnoor von Carolsfeld (1764–1841). In the Kupferstichkabinett, Dresden, Germany.

Marburg—Art Reference Bureau/EB Inc

Kant was cautious and still a long way from challenging the assumptions of Leibnizian metaphysics.

During the 15 years that he spent as a *Privatdozent,* Kant's renown as a teacher and writer steadily increased. Soon he was lecturing on many subjects other than physics and mathematics—including logic, metaphysics, and moral philosophy. He even lectured on fireworks and fortifications and every summer for 30 years taught a popular course on physical geography. He enjoyed great success as a lecturer; his lecturing style, which differed markedly from that of his books, was humorous and vivid, enlivened by many examples from his reading in English and French literature, and in travel and geography, science and philosophy.

Although he twice failed to obtain a professorship at Königsberg, he refused to accept offers that would have taken him elsewhere—including the professorship of poetry at Berlin that would have brought greater prestige. He preferred the peace and quiet of his native city in which to develop and mature his own philosophy.

Critic of Leibnizian Rationalism. During the 1760s he became increasingly critical of Leibnizianism. According to one of his students, Kant was then attacking Leibniz, Wolff, and Baumgarten, was a declared follower of Newton, and expressed great admiration for the moral philosophy of the Romanticist Jean-Jacques Rousseau.

His principal work of this period was *Untersuchung über die Deutlichkeit der Grundsätze der natürlichen Theologie und der Moral* (1764; "An Inquiry into the Distinctness of the Fundamental Principles of Natural Theology and Morals"). In this work he attacked the claim of Leibnizian philosophy that philosophy should model itself on mathematics and aim at constructing a chain of demonstrated truths based on self-evident premises. Kant argued that mathematics proceeds from definitions that are arbitrary, by means of operations that are clearly and sharply defined, upon concepts that can be exhibited in concrete form. In contrast with this method, he argued that philosophy must begin with concepts that are already given, "though confusedly or insufficiently determined," so that philosophers cannot begin with definitions without thereby shutting themselves up within a circle of words. Philosophy cannot, like mathematics, proceed synthetically; it must analyze and clarify. The importance of the moral order, which he had learned from Rousseau, reinforced the conviction received from his study of Newton that a synthetic philosophy is empty and false.

The attack on Leibniz

Besides attacking the methods of the Leibnizians, he also began criticizing their leading ideas. In an essay *Versuch, den Begriff der negativen Grössen in die Weltweisheit einzuführen* (1763), he argued that physical opposition as encountered in things cannot be reduced to logical contradiction, in which the same predicate is both affirmed and denied, and, hence, that it is pointless to reduce causality to the logical relation of antecedent and consequent. In an essay of the same year, *Der einzig mögliche Beweisgrund zu einer Demonstration des Daseyns Gottes,* he sharply criticized the Leibnizian concept of Being by charging that the so-called ontological argument, which would prove the existence of God by logic alone, is fallacious because it confuses existential with attributive statements: existence, he declared, is not a predicate of attribution. Moreover, with regard to the nature of space, Kant sided with Newton in his confrontation with Leibniz. Leibniz' view that space is "an order of co-existences" and that spatial differences can be stated in conceptual terms, he concluded to be untenable.

Investigation of the world of spirits

Some indication of a possible alternative of Kant's own to the Leibnizian position can be gathered from his curious *Träume eines Geistersehers erläutert durch Träume der Metaphysik* (1766). This work is an examination of the whole notion of a world of spirits, in the context of an inquiry into the spiritualist claims of Emanuel Swedenborg, a scientist and biblical scholar. Kant's position at first seems to have been completely skeptical, and the influence of the Scottish Skeptic David Hume is more apparent here than in any previous work; it was Hume, he later claimed, who first awoke him from his dogmatic slumbers. Yet Kant was not so much arguing that the notion of a world of spirits is illusory as insisting that men have no insight into the nature of such a world, a conclusion that has devastating implications for metaphysics as the Leibnizians conceived it. Metaphysicians can dream as well as spiritualists, but this is not to say that their dreams are necessarily empty; there are already hints that moral experience can give content to the ideal of an "intelligible world." Rousseau thus acted upon Kant here as a counterinfluence to Hume.

Early years of the professorship at Königsberg. Finally, in 1770, after serving for 15 years as a *Privatdozent,* Kant was appointed to the chair of logic and metaphysics, a position in which he remained active until a few years before his death. In this period—usually called his critical period, because in it he wrote his great *Critiques*—he published an astounding series of original works on a wide variety of topics, in which he elaborated and expounded his philosophy.

"Inaugural Dissertation" of 1770

The *Inaugural Dissertation* of 1770 that he delivered on assuming his new position already contained many of the important elements of his mature philosophy. As indicated in its title, *De Mundi Sensibilis atque Intelligibilis Forma et Principiis: Dissertatio,* the implicit dualism of the *Träume* is made explicit; and it is made so on the basis of a wholly un-Leibnizian interpretation of the distinction between sense and understanding. Sense is not, as Leibniz had supposed, a confused form of thinking but a source of knowledge in its own right, although the objects so known are still only "appearances"—the term that Leibniz also used. They are appearances because all sensing is conditioned by the presence, in sensibility, of the forms of time and space, which are not objective characteristics or frameworks of things but "pure intuitions." But though all knowledge of things sensible is thus of phenomena, it does not follow that nothing is known of things as they are in themselves. Certainly, man has no intuition, or direct insight, into an intelligible world; but the presence in him of certain "pure intellectual concepts, such as those of possibility, existence, necessity, substance, cause, enables him to have some descriptive knowledge of it. By means of these concepts he can arrive at an exemplar that provides him with "the common measure of all other things as far as real." This exemplar gives man an idea of perfection for both the theoretical and practical orders: in the first, it is that of the Supreme Being, God; in the latter, that of moral perfection.

After the *Dissertation,* Kant published virtually nothing for 11 years. Yet, in submitting the *Dissertation* to a friend at the time of its publication, he wrote:

> About a year since I attained that concept which I do not fear ever to be obliged to alter, though I may have to widen it, and by which all sorts of metaphysical questions can be tested in accordance with entirely safe and easy criteria, and a sure decision reached as to whether they are soluble or insoluble.

PERIOD OF THE THREE "CRITIQUES"

In 1781 the *Kritik der reinen Vernunft* (spelled "Critik" in the first edition; *Critique of Pure Reason*) was published, followed for the next nine years by great and original works that in a short time brought a revolution in philosophical thought and established the new direction in which it was to go in the years to come.

The Critique of Pure Reason. The *Critique of Pure Reason* was the result of some 10 years of thinking and meditation. Yet, even so, Kant published the first edition only reluctantly after many postponements; for although convinced of the truth of its doctrine, he was uncertain and doubtful about its exposition. His misgivings proved well-founded, and Kant complained that interpreters and critics of the work were badly misunderstanding it. To correct these wrong interpretations of his thought he wrote the *Prolegomena zu einer jeden künftigen Metaphysik die als Wissenschaft wird auftreten können* (1783) and brought out a second and revised edition of the first "critique" in 1787. Controversy still continues regarding the merits of the two editions: readers with a preference for an Idealistic interpretation usually prefer the first edition, whereas those with a Realistic view adhere to the second. But with regard to difficulty and ease of reading and understanding, it is generally agreed that there is little to choose between them. Anyone on first opening either book finds it overwhelmingly difficult and impenetrably obscure.

The cause for this difficulty can be traced in part to the works that Kant took as his models for philosophical writing. He was the first great modern philosopher to spend all of his time and efforts as a university professor of the subject. Regulations required that in all lecturing a certain set of books be used, with the result that all of Kant's teaching in philosophy had been based on such handbooks as those of Wolff and Baumgarten, which abounded in technical jargon, artificial and schematic divisions, and great claims to completeness. Following their example, Kant accordingly provided a highly artificial, rigid, and by no means immediately illuminating scaffolding for all three of his *Critiques*.

The *Critique of Pure Reason*, after an introduction, is divided into two parts, of very different lengths: A "Transcendental Doctrine of Elements," running to almost 400 pages in a typical edition, followed by a "Transcendental Doctrine of Method," which reaches scarcely 80 pages. The ". . . Elements" deals with the sources of human knowledge, whereas the ". . . Method" draws up a methodology for the use of "pure reason" and its a priori ideas. Both are "transcendental," in that they are presumed to analyze the roots of all knowledge and the conditions of all possible experience. The "Elements" is divided, in turn, into a "Transcendental Aesthetic," a "Transcendental Analytic," and a "Transcendental Dialectic."

The simplest way of describing the contents of the *Critique* is to say that it is a treatise about metaphysics: it seeks to show the impossibility of one sort of metaphysics and to lay the foundations for another. The Leibnizian metaphysics, the object of his attack, is criticized for assuming that the human mind can arrive, by pure thought, at truths about entities, which, by their very nature, can never be objects of experience, such as God, human freedom, and immortality. Kant maintained, however, that the mind has no such power and that the vaunted metaphysics is thus a sham.

The synthetic a priori in mathematics and physics

As Kant saw it, the problem of metaphysics, as indeed of any science, is to explain how, on the one hand, its principles can be necessary and universal (such being a condition for any knowledge that is scientific) and yet, on the other hand, involve also a knowledge of the real and so provide the investigator with the possibility of more knowledge than is analytically contained in what he already knows; *i.e.*, than is implicit in the meaning alone. To meet these two conditions, Kant maintained, knowledge must rest on judgments that are a priori, for it is only as they are separate from the contingencies of experience that they could be necessary and yet also synthetic; *i.e.*, so that the predicate term contains something more than is analytically contained in the subject. Thus, for example, the proposition that all bodies are extended is not synthetic but analytic because the notion of extension is contained in the very notion of body; whereas the proposition that all bodies are heavy is synthetic because weight supposes, in addition to the notion of body, that of bodies in relation to one another. Hence, the basic problem, as Kant formulated it, is to determine "How [*i.e.*, under what conditions] are synthetic a priori judgments possible?"

This problem arises, according to Kant, in three fields, viz., in mathematics, physics, and metaphysics; and the three main divisions of the first part of the *Critique* deal respectively with these. In the "Transcendental Aesthetic," Kant argued that mathematics necessarily deals with space and time and then claimed that these are both a priori forms of human sensibility that condition whatever is apprehended through the senses. In the "Transcendental Analytic," the most crucial as well as the most difficult part of the book, he maintained that physics is a priori and synthetic because in its ordering of experience it uses concepts of a special sort. These concepts—"categories," he called them—are not so much read out of experience as read into it and, hence, are a priori, or pure, as opposed to empirical. But they differ from empirical concepts in something more than their origin: their whole role in knowledge is different; for, whereas empirical concepts serve to correlate particular experiences and so to bring out in a detailed way how experience is ordered, the categories have the function of prescribing the general form that this detailed order must take. They belong, as it were, to the very framework of knowledge. But although they are indispensable for objective knowledge, the sole knowledge that the categories can yield is of objects of possible experience; they yield valid and real knowledge only when they are ordering what is given through sense in space and time.

The synthetic a priori in metaphysics

In the "Transcendental Dialectic" Kant turned to consideration of a priori synthetic judgments in metaphysics. Here, he claimed, the situation is just the reverse from what it was in mathematics and physics. Metaphysics cuts itself off from sense experience in attempting to go beyond it and, for this very reason, fails to attain a single true a priori synthetic judgment. To justify this claim, Kant analyzed the use that metaphysics makes of the concept of the unconditioned. Reason, according to Kant, seeks for the unconditioned or absolute in three distinct spheres: (1) in philosophical psychology it seeks for an absolute subject of knowledge; (2) in the sphere of cosmology, it seeks for an absolute beginning of things in time, for an absolute limit to them in space, and for an absolute limit to their divisibility; and (3) in the sphere of theology, it seeks for an absolute condition for all things. In each case, Kant claimed to show that the attempt is doomed to failure by leading to an antinomy in which equally good reasons can be given for both the affirmative and the negative position. The metaphysical "sciences" of rational psychology, rational cosmology, and natural theology, familiar to Kant from the text of Baumgarten, on which he had to comment in his lectures, thus turn out to be without foundation.

With this work, Kant proudly asserted that he had accomplished a Copernican revolution in philosophy. Just as the founder of modern astronomy, Nicolaus Copernicus, had explained the apparent movements of the stars by ascribing them partly to the movement of the observers, so Kant had accounted for the application of the mind's a priori principles to objects by demonstrating that the objects conform to the mind: in knowing, it is not the mind that conforms to things but instead things that conform to the mind.

The Critique of Practical Reason. Because of his insistence on the need for an empirical component in knowledge and his antipathy to speculative metaphysics, Kant is sometimes presented as a Positivist before his time; and his attack upon metaphysics was held by many in his own day to bring both religion and morality down with it. Such, however, was certainly far from Kant's intention. Not only did he propose to put metaphysics "on the sure path of science," he was prepared also to say that he "inevitably" believed in the existence of God and in a future life. It is also true that his original conception of his critical philosophy anticipated the preparation of a critique of moral philosophy. The *Kritik der praktischen Vernunft* (1788, spelled "Critik" and "practischen"; *Critique of Practical Reason*), the result of this intention, is the standard source book for his ethical doctrines. The earlier *Grundlegung zur Metaphysik der Sitten* (1785) is a shorter and, despite its title, more readily comprehensible treatment of the same general topic. Both differ from *Die Metaphysik der Sitten* (1797) in that they deal with pure ethics and try to elucidate basic principles; whereas the later work is concerned with applying what they establish in the concrete, a process that involved the consideration of virtues and vices and the foundations of law and politics.

Similarity between his ethics and epistemology

There are many points of similarity between Kant's ethics and his epistemology, or theory of knowledge. He used the same scaffolding for both—a "Doctrine of Elements," including an "Analytic" and a "Dialectic," followed by a "Methodology"; but the second *Critique* is far shorter and much less complicated. Just as the distinction between sense and intelligence was fundamental for the former, so is that between the inclinations and moral reason for the latter. And just as the nature of the human cognitive situation was elucidated in the first *Critique* by reference to the hypothetical notion of an intuitive understanding, so is that of the human moral situation clarified by reference to the notion of a "holy will." For a will of this kind there would be no distinction between reason and inclination; a being possessed of a holy will would always act as it ought. It would not, however, have the concepts of duty and moral obligation, which enter only when reason and desire find themselves opposed. In the case of human beings, the opposition is continuous, for man is at the same time both flesh and spirit; it is here that the influence of Kant's religious background is most prominent. Hence, the moral life is a continuing struggle in which morality appears to the potential delinquent in the form of a law that demands to be obeyed for its own sake—a law, however, the commands of which are not issued by some alien authority but represent the voice of reason, which the moral subject can recognize as his own.

In the "Dialectic," Kant took up again the ideas of God, freedom, and immortality. Dismissed in the first *Critique* as objects that men can never know because they transcend human sense experience, he now argued that they are essential postulates for the moral life. Though not reachable in metaphysics, they are absolutely essential for moral philosophy.

Kant is often described as an ethical Rationalist, and the description is not wholly inappropriate. He never espoused, however, the radical Rationalism of some of his contemporaries nor of more recent philosophers for whom reason is held to have direct insight into a world of values or the power to intuit the rightness of this or that moral principle. Thus, practical, like theoretical, reason was for him formal rather than material—a framework of formative principles rather than a content of actual rules. This is why he put such stress on his first formulation of the categorical imperative: "Act only on that maxim through which you can at the same time will that it should become a universal law." Lacking any insight into the moral realm, men can only ask themselves whether what they are proposing to do has the formal character of law—the character, namely, of being the same for all persons similarly circumstanced.

The Critique of Judgment. The *Kritik der Urteilskraft* (1790: spelled "Critik")—one of the most original and instructive of all of Kant's writings—was not foreseen in his original conception of the critical philosophy. Thus it is perhaps best regarded as a series of appendixes to the other two *Critiques*. The work falls into two main parts, called respectively "Critique of Aesthetic Judgment" and "Critique of Teleological Judgment." In the first of these, after an introduction in which he discussed "logical purposiveness," he analyzed the notion of "aesthetic purposiveness" in judgments that ascribe beauty to something. Such a judgment, according to him, unlike a mere expression of taste, lays claim to general validity; yet it cannot be said to be cognitive because it rests on feeling, not on argument. The explanation lies in the fact that, when a person contemplates an object and finds it beautiful, there is a certain harmony between his imagination and his understanding, of which he is aware from the immediate delight that he takes in the object. Imagination grasps the object and yet is not restricted to any definite concept; whereas a person imputes the delight that he feels to others because it springs from the free play of his cognitive faculties, which are the same in all men.

Critiques of aesthetics and natural teleology

In the second part, Kant turned to consider teleology in nature as it is posed by the existence in organic bodies of things of which the parts are reciprocally means and ends to each other. In dealing with these bodies, one cannot be content with merely mechanical principles. Yet if mechanism is abandoned and the notion of a purpose or end of nature is taken literally, this seems to imply that the things to which it applies must be the work of some supernatural designer; but this would mean a passing from the sensible to the suprasensible, a step proved in the first *Critique* to be impossible. Kant answered this objection by admitting that teleological language cannot be avoided in taking account of natural phenomena; but it must be understood as meaning only that organisms must be thought of "as if" they were the product of design, and that is by no means the same as saying that they are deliberately produced.

LAST YEARS

The critical philosophy was soon being taught in every important German-speaking university, and young men flocked to Königsberg as a shrine of philosophy. In some cases, the Prussian government even undertook the expense of their support. Kant came to be consulted as an oracle on all kinds of questions, including such subjects as the lawfulness of vaccination. Such homage did not interrupt Kant's regular habits. Scarcely five feet tall, with a deformed chest, and suffering from weak health, he maintained throughout his life a severe regimen. It was arranged with such regularity that people set their clocks according to his daily walk along the street named for him, "The Philosopher's Walk." Until old age prevented him, he is said to have missed this regular appearance only on the occasion when Rousseau's *Émile* so engrossed him that for several days he stayed at home.

With the publication of the third *Critique*, Kant's main philosophical work was done. From 1790 his health began to decline seriously. He still had many literary projects but found it impossible to write more than a few hours a day. The writings that he then completed consist partly of an elaboration of subjects not previously treated in any detail, partly of replies to criticisms and to the clarification of misunderstandings. With the publication in 1793 of his work *Die Religion innerhalb der Grenzen der blossen Vernunft*, Kant became involved in a dispute with Prussian authorities on the right to express religious opinions. The book was found to be altogether too Rationalistic for orthordox taste; he was charged with misusing his philosophy to the "distortion and depreciation of many leading and fundamental doctrines of sacred Scripture and Christianity" and was required by the government not to lecture or write anything further on religious subjects. Kant agreed but privately interpreted the ban as a personal promise to the King, from which he felt himself to be released on the latter's death in 1797. At any rate, he returned to the forbidden subject in his last major essay, *Der Streit der Fakultäten* (1798; "The Conflict of the Faculties").

Works on religion and the ban on them

The large work at which he laboured until his death—the fragments of which fill the two final volumes of the great Berlin edition of his works—was evidently intended to be a major contribution to his critical philosophy. What remains, however, is not so much an unfinished work as a series of notes for a work that was never written. Its original title was *Übergang von den metaphysische Anfangsgründe der Naturwissenschaft zur Physik* ("Transition from the Metaphysical Foundations of Natural Science to Physics"), and it may have been his intention to carry further the argument advanced in the *Metaphysische Anfangsgründe der Naturwissenschaft* (1786) by showing that it is possible to construct a priori not merely the general outline of a science of nature but a good many of its details as well. But judging from the extant fragments, however numerous they are, it remains conjectural whether its completion would have constituted a major addition to his philosophy and its reputation.

After a gradual decline that was painful to his friends as well as to himself, Kant died in Königsberg, February 12, 1804. His last words were "Es ist gut" ("It is good"). His tomb in the cathedral was inscribed with the words (in German) "The starry heavens above me and the moral law within me," the two things that he declared in the conclusion of the second *Critique* "fill the mind with ever new and increasing admiration and awe, the oftener and the more steadily we reflect on." (O.A.B.)

Kantianism

As a philosophical designation, Kantianism can signify either the system of thought contained in the writings of the epoch-making 18th-century philosopher Immanuel Kant or those later philosophies that arose from the study of Kant's writings and drew their inspiration from his principles. Only the latter is the concern of this section.

NATURE AND TYPES OF KANTIANISM

The Kantian movement comprises a loose assemblage of rather diverse philosophies that share Kant's concern with

Diverse trends in Kantianism

exploring the nature, and especially the limits, of human knowledge in the hope of raising philosophy to the level of a science in some sense similar to mathematics and physics. Participating in the critical spirit and method of Kant, these philosophies are thus opposed to dogmatism, to expansive speculative naturalism (such as that of Baruch Spinoza, the Jewish Rationalist), and, usually, to irrationalism. The various submovements of Kantianism are characterized by their sharing of certain "family resemblances"; *i.e.,* by the preoccupation of each with its own selection of concerns from among the many developments of Kant's philosophy: a concern, for example, with the nature of empirical knowledge; with the way in which the mind imposes its own categorial structure upon experience, and, in particular, with the nature of the structure that renders man's knowledge and moral action possible, a structure considered to be a priori (logically independent of experience); with the status of the *Ding an sich* ("thing-in-itself"), that more ultimate reality that presumably lurks behind man's apprehension of an object; or with the relationship between knowledge and morality. A brief exposition of Kant's philosophical system may be found above.

A system such as the critical philosophy of Kant freely lends itself to reconstructions of its synthesis according to whatever preferences the private philosophical inclinations of the reader may impose or suggest. Kant's system was a syncretism, or union, of British Empiricism (as in John Locke, George Berkeley, and David Hume) that stressed the role of experience in the rise of knowledge; of the scientific methodology of Isaac Newton; and of the metaphysical apriorism (or Rationalism) of Christian Wolff, who systematized the philosophy of Gottfried Leibniz, with its emphasis on mind. Thus it constituted a synthesis of elements very different in origin and nature, which tempted the student to read his own presuppositions into it.

Varieties of Kantianism

The critical philosophy has been subjected to a variety of approaches and methods of interpretation. These can be reduced to three fundamental types: those that conceive of the critical philosophy as an epistemology or a pure theory of (scientific) knowledge and methodology; those that conceive of it as a critical theory of metaphysics or the nature of Being (ultimate reality); and those that conceive of it as a theory of normative or valuational reflection parallel to that of ethics (in the field of action). Each of these types—known, respectively, as epistemological, metaphysical, and axiological Kantianism—can, in turn, be subdivided into several secondary approaches. Historically, epistemological Kantianism included such different attitudes as empirical Kantianism, rooted either in physiological or psychological inquiries; the logistic Kantianism of the Marburg school, which stressed essences and the use of logic; and the realistic Kantianism of the Austrian Alois Riehl. Metaphysical Kantianism developed from the transcendental Idealism of German Romanticism to Realism, a course followed by many speculative thinkers, who—like nearly all contemporary Kantians—saw in the critical philosophy the foundations of an essentially inductive metaphysics, in accordance with the results of the modern sciences. Finally, axiological Kantianism—concerned with value theory—branched, first, into an axiological approach (properly so-called), which interpreted the methods of all three of Kant's *Critiques* (*i.e., Critique of Pure Reason, Critique of Practical Reason,* and *Critique of Judgment*) as normative disciplines of thought; and, second, into an eclectic or relativistic Kantianism, which regarded the critical philosophy as a system of thought dependent upon social, cultural, and historical conditions.

The chief representatives of these submovements are identified in the historical sections below.

It is essential to distinguish clearly between two periods within the Kantian movement: first, the period from 1790 to 1831 (the death of Hegel); and, second, the period from 1860 to the present—separated by a time when an antiphilosophical Positivism, a type of thought that supplanted metaphysics with science, was predominant. The first period began with the thorough study and emendation of Kant's chief theoretical work, *Kritik der reinen Vernunft* (2nd ed., 1787; the *Critique of Pure Reason,* 1929); but it soon became intermingled with the romantic tendencies in German Idealism. The second period, called specifically Neo-Kantianism, was, first of all, a conscious reappraisal, in whole or in part, of the theoretical *Critique,* but also, as a total system, a reaction against Positivism. Earlier Neo-Kantianism reduced philosophy to the theory of knowledge and scientific methodology; systematic Neo-Kantianism, arising at the beginning of the 20th century, expressed itself in attempts at building metaphysical structures.

EARLY KANTIANISM: 1790–1835

Kant himself

According to Immanuel Kant, his major work, the *Critique of Pure Reason,* comprised a treatise on methodology, a preliminary investigation prerequisite to the study of science, which placed the Newtonian method (induction, inference, and generalization) over against that of Descartes and Wolff (deduction from intuitions asserted to be self-evident). The result was a critique of metaphysics, the value of which lay not in science but in a realm of being accessible only to the pure intellect. In exploring this "noumenal" realm, as he called it, Kant placed his *Critique* in a positive role. Recalling the revolution that occurred in astronomy when Nicolaus Copernicus discerned, in the apparent motions of the planets, reflections of the earth's own motion, Kant inaugurated a Copernican revolution in philosophy, which claimed that the subject doing the knowing constitutes, to a considerable extent, the object; *i.e.,* that knowledge is in part constituted by a priori or transcendental factors (contributed by the mind itself), which the mind imposes upon the data of experience. Far from being a description of an external reality, knowledge is, to Kant, the product of the knowing subject. When the data are those of sense experience, the transcendental (mental) apparatus constitutes man's experience or his science, or makes it to be such. These transcendental elements are of three different orders: at the lowest level are the forms of space and time (technically called intuitions); above these are the categories and principles of man's intelligence (among them substance, causality, and necessity); and at the uppermost level of abstraction the ideas of reason—the transcendental "I," the world as a whole, and God. It is by virtue of the encounter between the forms of man's sensory intuition (space and time) and his perceptions that phenomena are formed. The forms arise from the subject himself; the perceptions, however—or the data of experience—have reference, ultimately, to things-in-themselves, which nevertheless remain unknowable, inasmuch as, in order to be known at all, it is necessary for things to appear clothed, as it were, in the forms of man's intuition and, thenceforth, to present themselves as phenomena and not as noumena. The thing-in-itself, accordingly, indicates the limit and not the object of knowledge.

Early criticism

These theses of Kant provoked criticism among the followers of Christian Wolff, the Leibnizian Rationalist, and doubts among the disciples of Kant, which, as they further developed into systems, marked the first period of Kantianism. Inasmuch as these disciples took the *Critique of Pure Reason* to be a *preface* to the study of the pure reason or of the transcendental system and not the system itself, they saw in this interpretation an explanation for the ambiguities to which the *Critique* (as they felt) was subject. Their doubts revolved around two points: first, Kant had erroneously distinguished three kinds of a priori knowledge, coordinate with the three aforementioned levels or faculties of the mind; and second, Kant had accepted the thing-in-itself as constitutive of knowledge. Regarding the first point, they claimed that Kant had accepted the three faculties and their respective transcendental characteristics without investigation, in which case this structure should be viewed, in accordance with the preliminary character of the *Critique,* as a triple manifestation of a single fundamental faculty. For this reason the distinction between the levels of intuition and understanding (or between the receptivity and spontaneity of the mind) had to be rejected—for the three transcendentals—space and time, the categories, and the ideas of reason—were not existents but were only functions of thought. Finally, these disciples argued that the existence of a single transcendental subject, the Ego, would render the thing-in-itself superfluous and even pernicious for the scientific treatment of epistemology.

This function of human thought (the transcendental subject), which serves as the absolute source of the a priori, was variously designated by different early Kantian thinkers: for the German Realist Karl L. Reinhold, it constituted the faculty of representation; for the Lithuanian Idealist Salomon Maimon, it was a mental capacity for constructing objects; for the Idealist Jakob S. Beck, a protégé of Kant's, it was the act of synthesis; for the empirical critic of Kantianism G.E. Schulze, it was experience in the sense intended by Hume—a volley of discrete sense impressions; for the theory of knowledge of the outstanding ethical Idealist Johann G. Fichte, it was the original positing of the Ego and the non-Ego—which meant, in turn, in the case of the aesthetic Idealist F.W.J. von Schelling, "the absolute self," and in the case of G.W.F. Hegel "the *Geist* or absolute Spirit," and finally, in the case of the pessimistic Romanticist Arthur Schopenhauer, "the absolute Will." In each case (excepting Schulze) the interpretation of the thing-in-itself in a realistic metaphysical sense was rejected in favour of various degrees of transcendental Idealism. Removed from the main current of Kantianism was the empirically oriented thinker Jakob Friedrich Fries (the one figure in this group who was not an Idealist in the true sense), who interpreted the a priori in terms of psychological faculties and elements.

Having earlier renounced these apostates on a large scale, Kant, at the end of his life, prepared a new exposition of the transcendental philosophy (the second part of his *Opus Postumum*), which showed that he was ready tacitly to accede to the criticisms of his adversaries.

NEO-KANTIANISM: SINCE 1860

General features of Neo-Kantianism

Nineteenth-century Neo-Kantianism. The rejection of all of philosophy by Positivism had the anomalous effect of, itself, evoking an awakening of Kantianism, for many thinkers wished to give to Positivism itself a philosophical foundation that, while respecting the phenomenological attitude, would yet be hostile to the metaphysics of Positivism, which was usually a tacit, but inconsequent, Materialism. It was justifiably held that Kant could provide such a foundation because of his opposition to metaphysics and his limitation of scientific knowledge to the sphere of phenomena. The complexity of the critical philosophy was such that the theoretical criticism could be approached in diverse ways and that, through the facts themselves, diverse interpretations of the *Critique of Pure Reason* could be obtained. In the order of their origin (though not of their worth or importance), there thus arose currents of Kantianism that were empiricist, logicist, realist, metaphysical, axiological, and psychological—of which the most important have survived into the 20th century.

The return to Kant was determined by the historical fresco of the incomparable historian of philosophy Kuno Fischer entitled *Kants Leben und die Grundlagen seiner Lehre* (1860; "Kant's Life and the Foundations of his Teaching"), which replaced the earlier work of the semi-Kantian Ernst Reinhold, son of the more notable Jena scholar (published 1828–30), and especially that of the outstanding historian of philosophy Johann Eduard Erdmann (published 1834–53). In 1865 the order: "Zurück nach Kant!" ("Back to Kant!") reverberated through the celebrated work of the young epistemologist Otto Liebmann, *Kant und die Epigonen* ("Kant and his Followers"), which was destined to extricate their spirits from the Positivistic morass and, at the same time, to divert the Germans from romantic Idealism.

Schools of Neo-Kantianism

Epistemological Neo-Kantianism. The empiricist, logistic, and realistic schools can be classed as epistemological.

Empiricist Neo-Kantianism was represented by the erudite pioneering physicist and physiologist Hermann L.F. von Helmholtz and, in part, by F.A. Lange, author of a famous study of Materialism. Helmholtz found support in Kant for his claim, first, that, although perception can *represent* an external thing, it usually does so in a way far removed from an actual description of its properties; second, that space and time comprise an empirical framework created for thought by the perceiving subject; and, third, that causality is an a priori law allowing the philosopher to infer a reality that is absolutely unknowable. Similarly, Lange reduced science to the phenomenal level and repudiated the thing-in-itself.

Logistic Neo-Kantianism, as represented in the most well-known and flourishing school of Kantianism, that at Marburg, originated with Hermann Cohen, successor of Lange, who, in a book on Kant (1871), argued that the transcendental subject is not to be regarded as a psychic being but as a logical function of thought that constructs both the form *and* the content of knowledge. Nothing is *gegeben* ("given"), he urged; all is *aufgegeben* ("propounded," like a riddle) to thought—as when, in the infinitesimal calculus, the analyst generates motion by imagining thin slices of space and time and adding up their areas. Hence experience is a perfect construction of man's logical spirit. Cohen's example inspired many authors, among them Cohen's colleague at Marburg Paul Natorp, who, in his work on the logical foundations of the exact sciences, integrated even psychology into the Marburgian transcendentalism; and Ernst Cassirer, best known for stressing the symbolizing capacities of man, who, in his memorable work *Das Erkenntnisproblem in der Philosophie und Wissenschaft der neueren Zeit* (1906–20; *The Problem of Knowledge: Philosophy, Science, and History since Hegel,* 1966), transposed this same logisticism into a form that illumines the history of modern philosophy.

Realistic Neo-Kantianism, the third manifestation of epistemological Neo-Kantianism, was represented in the Realism of the scientific monist Alois Riehl and of his disciple Richard Hönigswald. In a work on the significance of the critical philosophy for the positive sciences (published 1876–87), Riehl held, in direct opposition to the Marburgian logisticism, that the thing-in-itself participates positively in the constitution of knowledge inasmuch as all perception includes a reference to things outside the subject.

Metaphysical Neo-Kantianism. Ten years after the appearance of the aforementioned ground-breaking book *Kant und die Epigonen,* its author, Otto Liebmann, introduced the new metaphysical approach in his book on the analysis of reality (1876), which came near to the Kantianism of Marburg. The Romanticist Johannes Volkelt, in turn, took up the theme of a critical metaphysics and expressed his persisting aspirations toward the Absolute in the claim that, beyond the certainties of man's own subjective consciousness, there exists a new kind of certainty in a transsubjective realm. Subjectivity is, thus, inevitably transcended, just as the sciences are surmounted when they presuppose a metaphysics. The influential spiritual moralist Friedrich Paulsen defended the claim that Kant had always behaved as a metaphysician, even in the *Critique of Pure Reason,* in spite of the epistemological restrictions that he imposed upon himself—a claim that made an impact that was felt throughout the following century.

Axiological Neo-Kantianism. Inasmuch as the two principal representatives of the axiological interpretation both taught at Heidelberg, this branch is also known as the Southwest German or Baden school. Its initiator was Wilhelm Windelband, esteemed for his "problems" approach to the history of philosophy. The scholar who systematized this position was his successor Heinrich Rickert, who had come from the tradition of Kuno Fischer. Drawing a parallel between the constraints that logic exerts upon thought and those that the sense of ought exerts upon ethical action, these thinkers argued that, while man's *action* must answer to an absolute value (the Good), his *thought* must answer to a regulative value (the True), which imposes upon him the duty of conforming to it. The *Critique of Pure Reason,* they held, elaborates this rule—which is not an entity but an imperative, or absolute, charge to act. Rickert regarded the critical endeavour as having been too narrow, since it was suited merely to physics. Actually, he charged, it should be the foundation for all of the sciences of the spirit. The distinctive characteristic of this school thus consisted in reintegrating German Idealism (as in Fichte and Hegel) into a rather personal Kantianism. Consequently, it succeeded in annexing more than one area of semi-Kantian thought: *e.g.*, "the philosophy of the spiritual sciences" of Wilhelm Dilthey, who held that intellectual life cannot be explained by means of naturalistic causality but only through historical understanding (*Verstehen*); "the life-philosophy" of the social philosopher Georg Simmel, who deviated from an earlier naturalistic relativism to the espousal of objective values; "the philosophy of value" of the experimental psychologist Hugo Münsterberg, author of one of the earliest systems of values; the "semi-Hegelianism" of Richard Kroner, a philosopher of culture and religion; and the general works of Bruno Bauch, Liebmann's successor at Jena. All of these philosophers were more or less related to axiological Neo-Kantianism.

Psychological Neo-Kantianism. An initial attempt to interpret Kantian transcendentalism in psychological terms was made by the Friesian Empiricist Jürgen Bona Meyer in his *Kants Psychologie* (1870). Later, a more important contribution in this field was made by the Göttingen philosopher of ethics and law Leonard Nelson and published in the *Abhandlungen der Fries'schen Schule* (1904 ff; "Acts of the Friesian School"). Even this title suggests an intimate agreement with the Kantianism of Fries's new critique of reason (1807); and Nelson, indeed, is regarded as the founder of the Neo-Friesian school. At a time when other Kantian schools were concerned with the transcendental analysis of objective or outer knowledge, Nelson held that, in the analysis of the subjective or inner self, the transcendental equipment of the mind—the a priori—is directly revealed. It thus fell to psychology to lay bare this equipment, which belongs in itself to the metaphysical order. It was upon this basis that the Marburg theologian Rudolf Otto, in his book *Das Heilige* (1917; *The Idea of the Holy,* 1958), ventured a type of religious phenomenology that has proved very successful.

Kantian philology

A discipline known as the Kant *Philologie,* concerned with the history, development, and works of Kant, has preempted a considerable portion of philosophical historiography since 1860. These studies began with the immense commentary on the *Critique of Pure Reason* produced in 1881–92 by Hans Vaihinger, known for his philosophy of the "As If" (which stresses man's reliance on pragmatic fictions), and with the founding of the new journal *Kantstudien* (1896) and the Kant-Gesellschaft ("Kantian Society," 1904)—both still extant. The most conspicuous result of this philological movement, however, was undeniably the monumental edition, in 23 volumes, of all of Kant's available works prepared (1900 ff) by the Academy of Sciences at Berlin under the editorship of the champion of humanistic studies, Wilhelm Dilthey. These volumes include: Sect. 1, Works; Sect. 2, Correspondence; Sect. 3, The "Nachlass." Since the transfer of this task to the University of Münster, Sect. 4, Kant's Lectures, has been undertaken. Those on logic and metaphysics (vols. 24–25) have been splendidly edited by Gerhard Lehmann.

Contemporary Neo-Kantianism. The recent development of Neo-Kantianism, except for innumerable historical studies, is very one-sided: no longer considered as exclusively epistemological, it merely prolongs the metaphysical school. Moreover, a large portion of the present Kant research is covered by the so-called *Problems of Kantianism* (see below). Important studies have been made on the development of Kant's philosophical thought, on Kant as a metaphysician, on his ontology and teachings on science, and on his transcendental deduction.

NON-GERMAN KANTIANISM

The Kantian awakening, in no wise limited to Germany, extended throughout Western philosophy. Its principal initiators were as follows: France was the first to open to its influence, beginning with the eclectic thinker Victor Cousin, who had studied German authors and made several trips to Germany. The relativistic personalist Charles Renouvier then defended a rather personal critical philosophy, which exerted an enduring influence through its impact upon the extreme Idealist Octave Hamelin of the Sorbonne; upon the metaphysician and cofounder of French neospiritualism Jules Lachelier; and upon his pupil, the philosopher of science Émile Boutroux.

English Kantianism

The English-speaking countries, on the other hand, have not seemed disposed to assimilate the critical philosophy as they did Hegelian Idealism. Except for the Scottish religious absolutist Edward Caird (*The Critical Philosophy of Immanuel Kant,* 1889), who was chiefly an Hegelian, there was in Britain at the close of the 19th century only another Scot, the critical Realist Robert Adamson, who was a Kantian. After him, however, can be cited the commentary, published in 1918, of Norman Kemp Smith, producer of the standard English translation of Kant's first *Critique,* and more recently, the remarkable exposition by the Oxford Kantian Herbert J. Paton, *Kant's Metaphysic of Experience* (2 volumes, 1936). Finally, Kantian methods can be discerned today in the later work of the prominent Oxford "ordinary language" philosopher, Peter F. Strawson, entitled *Individuals: An Essay in Descriptive Metaphysics* (1959). Kantianism became known in the United States toward 1840 primarily through the New England transcendentalist and poet Ralph Waldo Emerson—who was not, however, a Kantian himself. The physicist and logician Charles Sanders Peirce owes his Pragmatism largely to Kant's role as a counterweight against Hegelianism. The former southern California philosopher William H. Werkmeister represents a type of Neo-Kantianism inspired by the Marburg school (*The Basis and Structure of Knowledge,* 1948).

Italian scholars, on the other hand, have been vigorously engaged in Kantian studies since the initiative was taken by Alfonso Testa. The chief Neo-Kantian in Italy, however, was the Realist Carlo Cantoni, who took an anti-Positivist stance. Later, in the period from 1900 to 1918, Kantianism was represented by the extreme Realism of the theist Francesco Orestano. A school of Kantian philology has formed at Turin around the erudite Christian Idealist Augusto Guzzo and his journal *Filosofia.* More independent in spirit is the work of the critical ontologist Pantaleo Carabellese, Giovanni Gentile's successor at Rome.

ASSESSMENT OF KANTIANISM

At the present time there does not exist, either in Germany or elsewhere, a purely Kantian philosopher; but all acknowledge the obligation to come to grips with him. Within the four great currents of contemporary thought, however—*i.e.,* in Phenomenology, in the traditionalistic metaphysics, in Existentialism, and in the Positivistic Empiricism of the Vienna Circle and of Analytical philosophy—the predominant attitude toward Kant is negative.

Problems of Kantianism. As far as epistemology is concerned, the critical philosophy constitutes a theory of science that agrees with current trends; for science must have a base that is empirical though also real. On the other hand, the transcendental or a priori is implicated; and severe complications ensue whenever the question is posed whether a type of apprehension can be acquired apart from experience that conveys, however, some new and genuine knowledge—whether, in short, synthetic a priori judgments can be made. Significantly, the founder of Phenomenology, the German philosopher Edmund Husserl, came back to the fold of Kantian transcendentalism after previously opposing it bitterly. As against the Kantian position, Empiricism entirely rejects the possibility (and even the meaning) of the synthetic a priori and, robbed thereby of everything traditionally regarded as the subject matter of philosophy, directs its philosophical inquiries principally to the study of language. The foremost recent analyst of language, however, the pioneering philosopher Ludwig Wittgenstein, imposed upon philosophy the obligation to limit reason (or the transcendental element in knowledge)—a semi-Kantian position, which he nonetheless later renounced. As for Existentialism, one of recent Germany's foremost philosophers, Martin Heidegger, has presented in his *Kant und das Problem der Metaphysik* (1929; Eng. trans., *Kant and the Problem of Metaphysics,* 1962) a highly personalized interpretation. A student of Cohen at Marburg, the metaphysician Nicolai Hartmann, became the harbinger of the Realistic approach, elaborating in his analysis of the metaphysics of knowledge (1921) an ontological relation that he discerned to obtain between two forms of being: between thought and reality.

Empiricist criticism

Accordingly, the principles of thought correspond, in his view, to those of reality—a position at odds with Kant (even when he is interpreted as a Realist). Moreover, Hartmann treated the problems of mathematics—so urgent in current philosophy—in a manner that is again completely opposed to Kant; in particular, he questioned the validity of Kant's a priori intuition (or positing) of the spatiotemporal framework in terms of which man thinks about the world, challenging Kant at this point not merely to accommodate the non-Euclidean geometries (with curved space) that afforded a Realist alternative to the a priori but above all to reflect the distinctly logistic position regarding the foundations of mathematics to which he adhered. Discussion of the status of the thing-in-itself in man's knowledge of the real remained on the philosophical agenda both during and after Hartmann's time, but invoked the same indecision as it always had. At a time when Hartmann was accepting the thing-in-itself almost naïvely, Empiricism (in all its forms) rejected it categorically and attempted to construe the real in terms merely of what Kant had called phenomena. In the realm of ethics, Phenomenologists and Existentialists were dissatisfied with the purely formal character of Kant's ethics—*i.e.*, with its lack of specificity—and substituted a "material" ethic, of concrete duties, which was no less absolute than that of Kant. For their part, Empiricists were only interested in the analysis of expressions of moral judgment, which they reduced to imperative statements that are emotive and aimed at winning adherents.

Objections to Kantianism. It must be acknowledged that Kant has furnished many of the most significant themes that are found in the currents of contemporary philosophy, even in the forms that they still assume today. Yet, as compared with the state of affairs that existed from 1860 to 1918, Kantianism has suffered an impressive decline—though a slight recovery seems to have occurred during the third quarter of the 20th century.

Scientism, formalism, subjectivism

What were the reasons for this decline? In general, since World War I the reduction of philosophy to the philosophy of science has no longer been accepted, though contemporary Positivistic Empiricism has offered hardly any objection to it. The philosophy of science comprises, in fact, only *one* problem area, not the entire assemblage of philosophical problems. From this a second objection arises: Kantianism in general is too formalistic to satisfy man's actual inquisitiveness, which inclines more and more toward concrete concerns. Kantianism restricts itself to examining the a priori forms of thought and cares little for its diverse contents. Were this objection pertinent only to the exact sciences, it would not be serious—for these sciences attend to their own applications; but the objection becomes very grave for the field of ethics. For this reason, the objection against Kant's formalism has been raised most passionately against his ethical treatise, the *Critique of Practical Reason*—as by Hartmann, by the Phenomenologist Max Scheler, and by others. This transcendental formalism immediately encounters the further objection of subjectivism—in spite of efforts (from the side of logic) to evade it—*i.e.*, it is blamed for obstructing man's apprehension of the real universality of his Ego, of the thinking subject, and for inexorably impelling the scholar to the view that man's knowledge is merely the product of subjective construction. This subjectivistic transcendentalism, by its intrinsic logic, denies man access to the external world. Not only does it debar him from the world of things-in-themselves but it also prevents him from granting objective reality to phenomena as such, inasmuch as the transcendental source is here viewed as playing a constructive role with respect to experience and the phenomenon.

These three major objections, which stand out in the midst of many criticisms of minor details, recur constantly in the Kantian literature of the past quarter of a century. The result of these objections, as far as the evaluation of the critical philosophy is concerned, is that it is repudiated in its entirety—without, however, being thereby considered barred by limitation. Kant thus remains, in spite of everything, an inexhaustible source of problems and ideas, comparable in this respect to Plato and Aristotle, with whom he forms the great triad of Western philosophical thought. (H.J. de V.)

MAJOR WORKS

PRE-CRITICAL WRITINGS: *Gedanken von der wahren Schätzung der lebendigen Kräfte und Beurteilung der Beweise derer sich Herr von Leibniz und anderer Mechaniker in dieser Streitsache bedient haben* (1746); *Allgemeine Naturgeschichte und Theorie des Himmels* (1755; *Kant's Cosmogony . . .*, 1900 and 1968; *Universal Natural History and Theories of the Heavens*, 1969); *Principiorum Primorum Cognitionis Metaphysicae Nova Dilucidatio* (1755; Eng. trans. by F.E. England in *Kant's Conception of God*, 1929); *Metaphysicae cum geometria iunctae usus in philosophia naturali, cuius specimen I. continet Monadologiam physicam* (1756); *Versuch einiger Betrachtungen über den Optimismus* (1759); *Die falsche Spitzfindigkeit der vier syllogistischen Figurerewiesen* (1762; trans. in *Kant's Introduction to Logic and His Essay on the Mistaken Subtilty of the Four Figures*, 1963); *Der einzige mögliche Beweisgrund zu einer Demonstration des Daseyns Gottes* (1763; *Enquiry into the Proofs for the Existence of God*, 1836); *Versuch, den Begriff der negativen Grössen in die Weltweisheit einzuführen* (1763; *An Attempt to Introduce the Conception of Negative Quantities into Philosophy*, 1911); *Untersuchung über die Deutlichkeit der Grundsätze der natürlichen Theologie und der Moral* (1764); *Beobachtungen über das Gefühl des Schönen und Erhabenen* (1764, 1766, 1771; *Observations on the Feeling of the Beautiful and Sublime*, 1960); *Träume eines Geistersehers erläutert durch Träume der Metaphysik* (1766; *Dreams of a Spirit-Seer, Illustrated by Dreams of Metaphysics*, 1900; *Dreams of a Spirit Seer, and Other Related Writings*, 1969); *De Mundi Sensibilis atque Intelligibilis Forma et Principiis: Dissertatio* (1770; *Kant's Inaugural Dissertation and Early Writings on Space*, 1929); *Von den Verschiedenen Racen der Menschen* (1775).

CRITICAL AND POST-CRITICAL WRITINGS: *Critik der reinen Vernunft* (1781; rev. ed., *Kritik der reinen Vernunft*, 1787; *Critique of Pure Reason*, 1929, 1950); *Prolegomena zur einer jeden künftigen Metaphysik die als Wissenschaft wird auftreten können* (1783; *Prolegomena to Any Future Metaphysics*, 1951); *Grundlegung zur Metaphysik der Sitten* (1785; *The Fundamental Principles of the Metaphysic of Ethics*, 1938; *The Moral Law; or, Kant's Groundwork of the Metaphysic of Morals*, 1948; *Foundations of the Metaphysics of Morals*, 1969); *Metaphysische Anfangsgründe der Naturwissenschaft* (1786; *Metaphysical Foundations of Natural Science*, 1970); *Critik der practischen Vernunft* (1788; *Critique of Practical Reason*, 1949); *Critik der Urteilskraft* (1790, 2nd ed. 1793; *Kant's Kritik of Judgment*, 1892, reprinted as *Kant's Critique . . .*, 1914; new version, *Critique . . .*, vol. 1, *Kant's Critique of Aesthetic Judgment* and vol. 2, *Critique of Teleological Judgment*, 1911–28, republished 1952); *Über eine Entdeckung, nach der alle neu Critik der reinen Vernunft durch eine ältere entbehrlich gemacht werden soll* (1790; 2nd ed., 1791); *Die Religion innerhalb der Grenzen der blossen Vernunft* (1793; 2nd ed., 4 pt., 1794; *Religion Within the Boundary of Pure Reason*, 1838; *Religion Within the Limits of Reason Alone*, 2nd ed., 1960); *Zum ewigen Frieden* (1795; 2nd ed., 1796; *Project for a Perpetual Peace*, 1796, many later editions called *Perpetual Peace;* 1915 ed. reprinted 1972); *Die Metaphysik der Sitten* (1797; 2nd ed., 1798–1803; *The Metaphysic of Morals*, 2 vol., 1799 and 1965; *The Metaphysic of Ethics*, 1836), comprising *Metaphysische Anfangsgründe der Rechtslehre* (*The Philosophy of Law*, 1887) and *Metaphysische Anfangsgründe der Tugendlehre* (*The Doctrine of Virtue*, 1964); *Der Streit der Facultäten* (1798); *Von der Macht des Gemüths durch den blossen Vorsatz seiner krankhaften Gefühle Meister zu seyn* (1798; *Kant on the Art of Preventing Diseases*, 1806); *Anthropologie in pragmatischer Hinsicht abgefasst* (1798; improved ed., 1800; *The Classification of Mental Disorders*, 1964); *Immanuel Kants Physische Geographie*, 3 vol. in 6 pt., 1801–04); *I. Kants Logik: Ein Handbuch zu Vorlesungen* (1800; *Logic*, 1819); *Immanuel Kant über Pädagogik* (1803; *Kant on Education*, 1899; *The Educational Theory of Immanuel Kant*, 1904; *Education*, 1960); *Welches sind die wirklichen Fortschritte, die Metaphysik seit Leibnizens und Wolf's Zeiten in Deutschland gemacht hat?* (1804).

BIBLIOGRAPHY

Kant. *Biography:* The main sources for Kant's life are three memoirs published in 1804: LUDWIG ERNEST VON BOROWSKI, *Darstellung des Lebens und Charakters Immanuel Kants* (reprinted 1968); REINHOLD B. JACHMANN, *Immanuel Kant geschildert in Briefen an einen Freund* (reprinted 1968); and CHRISTOPH WASIANSKI, *Immanuel Kant in seinen letzten Lebensjahren* (the basis of THOMAS DE QUINCEY's "The Last Days of Kant" included in his *Works*). See also JOHN H.W. STUCKENBERG, *The Life of Immanuel Kant* (1882); FRIEDRICH PAULSEN, *Immanuel Kant: His Life and Doctrine* (1902, reissued 1963; originally published in German, 1898); ERNST CASSIRER, *Kant's Life and Thought* (1981; trans. of 2nd German ed., 1921);

and KARL VORLÄNDER, *Immanuel Kants Leben*, 3rd ed. (1974), and *Immanuel Kant: Der Mann und das Werk*, 2 vol. (1924, reissued 1977).

Editions: The standard edition of Kant's works is that of the Berlin Academy (later the DDR Academy), *Gesammelte Schriften* (1902–), 29 vol. by 1980, which contains Kant's lectures, correspondence, and literary remains as well as his published writings. There are also modern collected editions by ERNST CASSIRER, 11 vol. (1912–23); and by KARL VORLÄNDER, 10 vol. (1920–29). A convenient edition of the *Critique of Pure Reason* is that by RAYMOND SCHMIDT, 1926).

Aids to study: RUDOLF EISLER, *Kant-lexicon* (1930, reprinted 1971); HEINRICH RATKE, *Systematisches Handlexikon zu Kants Kritik der reinen Vernunft* (1929, reprinted 1965).

General works: The best introduction is STEPHAN KÖRNER, *Kant* (1955, reissued 1982). See also EDWARD CAIRD, *The Critical Philosophy of Immanuel Kant*, 2 vol. (1889, reprinted 1969); KUNO FISCHER, *Kants Leben und die Grundlagen seiner Lehre* (1860); ALOIS RIEHL, *Der philosophische Kriticismus . . .*, 3rd ed., 3 vol. (1924–26); BRUNO BAUCH, *Immanuel Kant*, 4th ed. (1933), in German; HENRICH RICKERT, *Kant als Philosoph der modernen Kultur* (1924); MAX WUNDT, *Kant als Metaphysiker* (1924); MARTIN HEIDEGGER, *Kant and the Problem of Metaphysics* (1962, reissued 1972; originally published in German, 1929); HERMAN J. DE VLEESCHAUWER, *La Déduction transcendentale dans l'oeuvre de Kant*, 3 vol. (1934, reissued 1976), and *L'Évolution de la pensée kantienne* (1939); PANTALEO CARABELLESE, *Il problema della filosofia in Kant* (1938); GOTTFRIED MARTIN, *Kant's Metaphysics and Theory of Science* (1955, reissued 1974; originally published in German, 1951); HEINZ HEIMSOETH, *Studien zur Philosophie Immanuel Kants*, 2nd ed. (1971); RICHARD KRONER, *Kant's Weltanschauung* (1956; originally published in German, 1914); FRIEDRICH DELEKAT, *Immanuel Kant*, 3rd ed. (1969), in German.

Precritical writings: MARIANO CAMPO, *La genesi del criticismo kantiano* (1953); GIORGIO TONELLI, *Elementi metodologici e metafisici in Kant dal 1745 al 1768* (1969), in German.

The "Critique of Pure Reason": NORMAN KEMP SMITH, *A Commentary to Kant's "Critique of Pure Reason,"* 2nd ed. rev. (1923, reissued 1979); HERBERT J. PATON, *Kant's Metaphysic of Experience*, 2 vol. (1936); ALFRED C. EWING, *A Short Commentary on Kant's Critique of Pure Reason* (1938, reprinted 1974); THOMAS D. WELDON, *Kant's Critique of Pure Reason*, 2nd ed. (1958); HERMANN COHEN, *Kants Theorie der Erfahrung*, 4th ed. (1925), and *Kommentar zu Immanuel Kants Kritik der reinen Vernunft*, 4th ed. (1925); HANS VAIHINGER, *Kommentar zu Kants Kritik der reinen Vernunft*, 2nd ed., 2 vol. (1922, reissued 1970); HEINZ HEIMSOETH, *Transzendentale Dialektik*, 3rd vol. (1966–69). See also GRAHAM BIRD, *Kant's Theory of Knowledge* (1962, reissued 1973); PETER F. STRAWSON, *The Bounds of Sense* (1966, reissued 1975).

Ethical writings: JEFFIRIE G. MURPHY, *Kant: The Philosophy of Right* (1970); PAUL A. SCHILPP, *Kant's Pre-Critical Ethics*, 2nd ed. (1960, reprinted 1977); HERBERT J. PATON, *The Categorical Imperative* (1947, reissued 1971); ORNA NELL, *Acting on Principle: An Essay on Kantian Ethics* (1975); VIGGO ROSSVAER, *Kant's Moral Philosophy: An Interpretation of the Categorical Imperative* (1979); LEWIS W. BECK, *A Commentary on Kant's Critique of Practical Reason* (1960); A.E. TEALE, *Kantian Ethics* (1951, reprinted 1975); WILLIAM D. ROSS, *Kant's Ethical Theory* (1954, reprinted 1978); HERMANN COHEN, *Kants Begründung der Ethik*, 2nd ed. (1910); MAX SCHELER, *Der Formalismus in der Ethik und die materiale Wertethik*, 6th ed. (1980). See also PAUL MENZER's ed. of *Eine Vorlesung Kants über Ethik* (1924; Eng. trans., *Lectures on Ethics by Immanuel Kant*, (1930); MORRIS STOCKHAMMER, *Kants Zurechnungsidee und Freitheitsantinomie* (1961); HENRICH W. CASSIRER, *A Commentary on Kant's Critique of Judgement* (1938, reprinted 1970); ALFRED BAEUMLER, *Kants Kritik der Urteilskraft* (1923).

Particular problems: (*Science*): ERICH ADICKES, *Kant als Naturforscher*, 2 vol. (1924–25); JULES VUILLEMIN, *Physique et métaphysique kantiennes* (1955). (*Ontology*): CHRISTOPHER B. GARNETT, *The Kantian Philosophy of Space* (1939, reprinted 1965); MARTIN HEIDEGGER, *Kants These über das Sein* (1963), and *What Is a Thing?* (1968; originally published in German, 1962). (*Philosophy of history*): YIRMIAHU YOVEL, *Kant and the Philosophy of History* (1980); KLAUS WEYAND, *Kants Beschichtsphilosophie: Ihre Entwicklung und ihr Verhältnis zur Aufklärung* (1963). (*Political philosophy*): SUSAN M. SHELL, *The Rights of Reason: A Study of Kant's Philosophy and Politics* (1980); GEORGES VLACHOS, *La Pensée politique de Kant* (1962). (*Religion*): CLEMENT C.J. WEBB, *Kant's Philosophy of Religion* (1926, reprinted 1970); JOSEF BOHATEC, *Die Religionsphilosophie Kants . . .* (1938, reprinted 1966); ALLAN W. WOOD, *Kant's Rational Theology* (1978). (*Comparative studies*): JOHANNES B. LOTZ (ed.), *Kant und die Scholastik heute* (1955); KARL JASPERS, *Die grossen Philosophen*, 2 vol. (1957–81; abridged Eng. trans., *The Great Philosophers*, ed. by HANNAH ARENDT, 1966). (*Aesthetics*): DONALD W. CRAWFORD, *Kant's Aesthetic Theory* (1974); PAUL GUYER, *Kant and the Claims of Taste* (1979).

Opus postumum: ERICH ADICKES, *Kants Opus Postumum* (1920, reprinted 1978); GERHARD LEHMANN, *Kants Nachiasswerk und die Kritik der Urteilskraft* (1939).

Journals: Many important works of Kantian scholarship have been published in the periodical *Kant-studien* (quarterly).

Kantianism. Though the literature on Kant himself comprises innumerable titles, that on Kantianism is relatively scanty. One work that contains the complete history of Kantianism is *Freidrich Ueberwegs Grundriss der Geschichte der Philosophie*, 13th ed. (1953), vol. 3:606–620 and 4:1–128 for the first period and pp. 410–483 for the second. For the first period, there is an abundant literature. Of particular interest for the present purposes are JOHANN E. ERDMANN, *Versuch einer wissenschaftlichen Darstellung der Geschichte der neuern Zeit*, 2nd ed., vol. 3 (1923); and G. LEHMANN, "Kant im Spätidealismus und die Anfänge der neukantischen Bewegung," in *Zeitschrift für philosophische Forschung*, 17:438–456 (1963). For the second period, MARIANO CAMPO has begun the history in his *Schizzo storico della esegesi e critica Kantiana* (1959); and summaries have been written by LEWIS W. BECK in the *Encyclopedia of Philosophy*, 5:468–473 (1967, reissued 1972); and by HERMANN NOACK in his *Die Philosophie Westeuropas*, pp. 143–196 (1962). See also the *Enciclopedia Filosofica*, new ed., vol. 3, col. 1225, and vol. 4, col. 953 (1967); WOLFGANG RITZEL, *Studien zum Wandel der Dantauffassung* (1952); HENRI DUSSORT, *L'École de Marbourg* (1963); and HEINRICH RICKERT, *Die Heidelberger Tradition und Kants Kritizismus* (1934).

Karāchi

Karāchi, the principal seaport and the largest city in Pakistan, is located on the coast of the Arabian Sea immediately northwest of the Indus River Delta. It is the capital of the province of Sind as well as the headquarters of the district of Karāchi. It is also a major commercial and industrial centre. The city proper covers an area of 228 square miles (591 square kilometres), while the metropolitan area of Greater Karāchi spreads out over an area of 560 square miles.

The city has been variously called Caranjee, Crochey, Krotchey, Currachee, and Kurrachee. All its names are believed to be derived from the Sindhi name of the original settlement that initially stood on the spot—Kalachi-Jo-goth (meaning the village of Kalachi—the headman of the tribe).

The impetus to Karāchi's development originally came from its role as the port serving the Indus River Valley and the Punjab region of British India. The development of air travel subsequently increased Karāchi's importance. It is also the port serving the landlocked country of Afghanistan.

This article is divided into the following sections:

Physical and human geography

THE LANDSCAPE

The city site. Karāchi Harbour, on the shores of which the city is situated, is a safe and beautiful natural harbour. It is protected from storms by Kiamāri Island, Manora Island, and Oyster Rocks, which together block the greater part of the harbour entrance in the west.

A low-lying coastal strip runs along the shore of the harbour. Away from the coast, the ground rises gently to the north and east to form a large plain, from five to 120 feet (1½ to 37 metres) above sea level, on which the city of Karāchi is built. The Malīr River, a seasonal stream, passes through the eastern part of the city, and the Layāri River, also seasonal, runs through the most densely populated northern section. Some ridges and isolated hills occur in the north and east; Mango Pīr, the highest elevation, is 585 feet high.

The 560 square miles that comprised the Federal Capital Area of Pakistan in 1948 are considered, for all practical purposes, to form the Karāchi metropolitan area. Almost half of the area is occupied by the city and its suburbs, and the surrounding 332 square miles consist of agricultural land and government wasteland.

Climate. Karāchi has pleasant weather for the greater part of the year. May and June are the hottest months, when the mean maximum temperature is about 93° F (34° C). Spells of enervating weather occasionally prevail in May and October, during which the temperature shoots up to 105° F (41° C). The coolest months are January and February, during which the mean minimum temperature remains about 56° F (13° C). A biting north wind occasionally blows in these months, during which the temperature may drop to 40° F (4° C). The relative humidity varies from 58 percent in October, the driest month, to 82 percent in August, the wettest month. The average rainfall is eight inches (203 millimetres); most of the rain falls during a total of nine or 10 days in the months of June, July, and August.

The city faces some pollution problems. High humidity in the region does not permit evaporation of stagnant water in some places, while fumes from factories and automobiles contribute to air pollution, in spite of land and sea breezes.

Plant and animal life. The natural vegetation is scanty. Seaweed rises in tangles, and mangroves grow along some of the shores. Coarse grass, cactus, and castor plants occur on the plains and hills, and date and coconut palms grow in the river valleys.

The common wild animals are wolves, chinkaras (a type of gazelle), hog deer, jackals, wild cats, and hares. Domestic animals include sheep, goats, horses, and cows. Local birds include geese, ducks, snipe (game birds related to the woodcock), cranes, flamingos, and ibis (wading birds related to the heron). Various types of snakes are found in the region, particularly cobras, kraits, vipers, and python.

The city layout. The most striking aspect of Karāchi's layout is the west-to-east parallel alignment of the four arterial roads—Nishter Road (formerly called Lawrence Road), Mohammed Ali Jinnah Road (formerly Bandar Road), Shahrah-e-Liaquat (Frere Road), and I.I. Chundrigar Road (McCleod Road). Beginning at Mereweather Tower in the vicinity of the port, these roads run through the centre of the city. Several roads, such as Napier Road, Dr. Zia-ud-din Ahmed Road (Kutchery Road), and Garden Road, cut perpendicularly across these arteries from north to south.

The old town

The old town lies near the port, to the north of M.A. Jinnah Road, and with extensions stretching along the material roads for over a mile; unplanned, it is reminiscent of medieval towns of the Near East or Europe. East of the old town are such districts as the Drigh Cantonment, the Civil Lines (residential areas for senior civil service officers), and the Saddar Bazar. This area is planned on a checkerboard pattern, and shows European characteristics. Beyond this stretch several radial roads, along which growth has taken the form of neighbourhood units; each unit is laid out with straight, broad roads connected by smaller streets.

Land use

The land-use pattern of the city is complex. In the central area, the preponderance of residential property tends to form a matrix within which all other functions are distributed. There is, however, a marked concentration of commercial buildings at the western ends of M.A. Jinnah Road and I.I. Chundrigar Road. Wholesale businesses are located in the old town, retail businesses along M.A. Jinnah Road and in Saddar Bazar, and the government offices on Shahrah-e-Liaquat, near Saddar. The outer areas are dominated by dormitory suburbs interspersed with a scattering of cantonments (military quarters), agricultural tracts, salt works, airports, railway stations, and marshalling yards.

The city proper has old and decayed buildings, occupied by members of the middle and lower income groups. Further from the city centre are modern bungalows occupied by richer persons; the outer zone is occupied by workers.

Karāchi has a variety of types of buildings. The central area contains apartment bungalows, barracks, and mul-

The city of Karāchi and (inset) its metropolitan area.

tistoried buildings; the outer areas are characterized by bungalows, blocks of flats, and quarters (streets of small houses). Buildings of the British period were constructed with stone in western styles of architecture; other stone buildings in the central city show a blending of Eastern and Western styles, and have towers, domes, pillars, arches, hanging balconies, and rectangular courtyards. Buildings in the outer areas are built of cement blocks, and with few exceptions they show no uniformity in design. Some follow contemporary North American design, while others incorporate features of traditional Muslim architecture.

THE PEOPLE

No ethnic group predominates in the city. Cultural and social activities essentially revolve around religion. The population is almost entirely Muslim, but there are also small Christian, Hindu, Parsi, Buddhist, and Jain minorities. Most of the Muslims derive from Indo-Pakistani stock, except for "Makranis" and "Shiddies," who are also descended from Negro elements, and who originated during the era of the slave trade in the days before British rule, when Karāchi was an important slave-trading centre. Some of the members of the Christian minority are of Indo-Pakistani origin, while others are descended from Portuguese or other European groups.

THE ECONOMY

Industry. Textiles and footwear are the principal items manufactured, followed by such items as metal products, food and beverages, paper and printing, wood and furniture, machinery, chemicals and petroleum, leather and rubber, and electrical goods. Karāchi is also an important centre for handicrafts and cottage industries that produce handloomed cloth, lace, carpets, articles made of brass and

Manufacturing

The Defence Society Mosque in Karāchi.
William MacQuitty—Camera Press

bell metal (an alloy of copper and tin), pottery, leather goods, and gold and silver embroidery. Karāchi handles the entire seaborne trade of Pakistan and of landlocked Afghanistan.

Finance. There are more than 25 banks in Karāchi that have branches throughout Pakistan; these include the State Bank of Pakistan, the Habib Bank Ltd., the National Bank of Pakistan, the United Bank Ltd., the Industrial Development Bank of Pakistan, and the Agricultural Development Bank of Pakistan. The city is also the centre of about two dozen insurance companies, which play an important role in the economic development of the country by investing large sums in power development, housing programs, jointstock companies, government loan securities, and savings certificates.

Karāchi has a stock exchange that handles an overwhelming proportion of transactions in government securities and in the shares of most of the important industrial and financial institutions.

Transportation. The Karāchi–Peshāwar highway links the city with the interior of Pakistan, while the Karāchi–Ormāra highway extends along the coast. The Karāchi to Zāhedān highway connects it with Iran and other Middle Eastern countries. Express roads radiate from the city centre, while feeder roads connect the express roads with local streets.

Karāchi is the terminus of Pakistan's railway system, which mainly serves to transport goods between Karāchi and the interior. There are also passenger trains, as well as a circular railway that skirts the city on the north and the east, for commuter traffic and the transport of goods between the port and the industrial areas.

Karāchi Airport provides international and domestic services. The port of Karāchi is one of the busiest east of Suez.

ADMINISTRATIVE AND SOCIAL CONDITIONS

Government. The city is administered by five institutions, the heads of which are appointed by the government. The Karāchi Municipal Corporation, constituted in 1852, performs a large number of civic functions affecting more than three-fourths of the population of Greater Karāchi. The Korangi-Lāndhi and Drigh-Malīr municipal committees were established in 1966 and 1970, respectively, to provide civic facilities to the suburban areas developed after 1947. The Karāchi Cantonment Board is the administrative body for the areas where the military are quartered. The Karāchi Port Trust administers the affairs of the port and is entrusted with the development and maintenance of the harbour.

Public utilities. The three main sources of the city's water supply are Lake Hāleji, 55 miles (90 kilometres) away, fed by the Indus River; wells that have been sunk in the dry bed of the Malīr River, 18 miles away; and Lake Kalri, 60 miles away, also fed by the Indus waters. Although the city's water mains stretch for many miles, some of the outer areas, such as Lāndhi, Malīr, New Karāchi, and Mauripur, still have an acute water shortage.

Sources of water supply

The Karāchi Electric Supply Corporation is responsible for electricity services. It has several power stations located in the city; these stations use natural gas, diesel oil, or both. A nuclear power station is operated at Paradise Point.

Karāchi Municipal Corporation maintains a fleet of vehicles for refuse collection, night soil removal, dog catching, and antimalarial and antifly operations. Sweepers are employed to clean the streets. Sewage is disposed of by two underground drainage systems, and there are two sewage treatment plants, one serving the city proper and the other the outlying areas.

Health and security. Karāchi proper has more than 20 general hospitals, as well as several hospitals specializing in tuberculosis, skin diseases, leprosy, and epidemic diseases. There are also child-welfare centres and dispensaries, in addition to general hospitals in the suburbs.

There are several well-equipped fire-fighting stations; separate fire brigade units are attached to the railway network. In addition, the Port Trust and Pakistan International Airlines (PIA) have services that can be used in emergencies.

The police are administered by the Sind provincial administration; the inspector general of police is assisted by a force of more than 1,100. The city is divided into 40 police districts.

Education. Karāchi has more than 900 schools, of which the majority are primary schools and the rest are secondary schools. More than half of all these are privately run, the rest being run by the government. Among schools established by different religious communities are Karāchi Grammar School, St. Joseph's Convent School, and St. Patrick's High School, all of which are Christian; a school for the Parsi community; and Sind Madrassa, a Muslim school.

The University of Karāchi

The University of Karāchi is the primary institution of higher education. It has more than 20 graduate departments in arts and sciences, as well as a graduate school of business administration. Courses in a variety of subjects, including commerce and law, are provided by about 75 colleges affiliated to the university. In addition, there is a medical college, as well as two engineering colleges, a polytechnic institute, a college of home economics, and two teacher-training colleges.

CULTURAL LIFE

The Arts Council of Pakistan is the primary cultural institution in the city; it organizes various cultural functions including art exhibitions and offers training in music. The Ghanshyam Art Centre and the Bulbul Academy promote Pakistani dancing and other cultural activities.

Karāchi does not have well-established theatre, but amateur dramas and variety shows are frequently staged in Katrak Hall. Motion pictures are more popular, and there are more than 50 cinemas.

Karāchi has a small museum containing relics of the early Indus Valley civilization and examples of the Greco-Buddhist art of Gandhāra (a region of ancient India in what is now northwestern Pakistan); it also has some ethnological collections representing life in different regions of Pakistan.

The library of the University of Karāchi is the city's largest, but there are other libraries containing books of a popular nature. Material of a more scholarly nature is to be found in the British Council Library, the American Center Library, and the Liaquat Memorial Library. The departmental libraries of the State Bank of Pakistan, the Pakistan Institute of Development Economics, and the National Archives contain collections of books on economics and on national matters.

There is a general shortage of open spaces and parks in Karāchi. Gandhi Gardens and Fatima Jinnah (Burns) Gardens are popular parks. There are a number of fine swimming and fishing beaches, such as Paradise Point, Hawkes Bay, Sandspit, Manora, and Clifton. The Karāchi Zoo is located in the Gandhi Gardens and contains a varied collection of mammals, birds, and reptiles.

Sports and games facilities are mostly provided by such associations as the Karāchi Gymkhana, the Parsi Gymkhana, the Agha Khan Gymkhana, and the Young Men's Christian Association (YMCA). Various organizations and educational institutions have their own playgrounds. The largest sports area is the National Stadium, which contains playgrounds for cricket, hockey, association football (soccer), and tennis. There are also boating, yachting, and flying clubs. (Z.A.K.)

History

Origins of the port

Karāchi was a small fishing village when a group of traders moved there in the early 18th century from the decaying port of Kharak Bandar nearby. Besides the natural protection against monsoon storms, Manora Head furnished an excellent site for the defense of the harbour, and the Talpura *amīrs* who gained Karāchi from the *khān* of Kalāt in 1795 erected a permanent fort on it. The settlement expanded rapidly and was already of significance when it was captured in 1839 by the British, who annexed it in 1842, together with the province of Sind. It then became an army headquarters for the British and also began to develop from a fishing village into the principal port for the Indus River region.

In 1843 a river-steamer service was introduced between Karāchi and Multān, about 500 miles up the Indus. Port facilities were improved from 1854 onward. In 1861 a railway was built from Karāchi to Kotri, 90 miles upstream on the right bank of the Indus, opposite Hyderābād. In 1864 direct telegraph communications were established with London and with the interior. With the opening of the Suez Canal in 1869, the importance of Karāchi grew, and it became a full-fledged seaport. By 1873 it possessed an efficient and well-managed harbour.

Karāchi was connected directly with the hinterland when the railway line was extended from Kotri in 1878 to join the Delhi-Punjab railway system at Multān. In 1886, the Karāchi Port Trust was established as the port authority, and between 1888 and 1910 the East Wharf—186,000 feet in length—was constructed. When the Punjab emerged as the granary of India in the 1890s, Karāchi became the region's principal outlet. By 1914 it had become the largest grain exporting port of the British Empire.

After World War I, manufacturing and service industries were installed. By 1924 an aerodrome had been built, and Karāchi became the main airport of entry to India. The city became the provincial capital of Sind in 1936.

Growth after 1947

With the creation of Pakistan in 1947, Karāchi not only became the capital and premier port of the new country but also a centre for industry, business, and administration. Although Rāwalpindi became the interim capital in 1959, before the capital permanently moved to Islāmābād in 1969, Karāchi maintained its preeminence as Pakistan's business and industrial hub. In a development typical of many postcolonial megacities, this premier economic status subsequently contributed to an enormous increase in Karāchi's population, as a huge influx of job-seeking immigrants from rural areas nearly doubled the city's size in the final two decades of the 20th century. With the city's infrastructure already overburdened to the breaking point, fully one-third of these new arrivals were forced to take up residence in urban shantytowns known as *katchi abadis*, which ordinarily lacked power, running water, or sanitation. The delivery of basic city services remained an ongoing problem for Karāchi into the 21st century.

The final quarter of the 20th century also brought a huge wave of urban violence and crime to Karāchi, in the form of ethnic violence between native Sindhis and more recent immigrants from India, the *muhajirs*, and in an increased rate of both simple crime and organized brigandage. The severity of lawlessness in Karāchi prompted the government to initiate a military crackdown on crime in the city in 1992, but this campaign only began to see significant results by the end of that decade.

BIBLIOGRAPHY. NAZIR AHMAD, *Survey of Shelterless Persons in Karachi 1959* (1959), is a detailed report of the problems faced by the city due to the influx of refugees from India following the establishment of Pakistan. J.J. VAN DER LINDEN, *Karachi: Migrants, Housing and Housing Policy* (1991), analyzes the housing requirements, activities, and needs in metropolitan Karāchi. A.F. BAILLIE, *Kurrachee: Past, Present and Future* (1890), gives Karāchi's history and its growth in the early years of British rule. R.F. BURTON, *Scinde; or, the Unhappy Valley*, 2 vol. (1851), and *Sind Revisited*, 2 vol. (1877), are classic surveys of Sind under British rule. *Population Growth and Policies in Mega-Cities: Karachi* (1988), population policy paper no. 13, published by the United Nations, is useful. H. FELDMAN, *Karachi Through a Hundred Years* (1960), written to commemorate the centenary of the Karāchi Chamber of Commerce and Industry, gives details of the role played by the Chamber in the economic development of Karāchi from 1860 to 1960. H.T. LAMBRICK, *Sind: A General Introduction*, 2nd ed. (1975), includes a discussion of Karāchi. RICHARD F. NYROP *et al.*, *Area Handbook for Pakistan* (1975), discusses many aspects of modern Karāchi. Karāchi Development Authority MP Reports have been written for the preparation of the Master Plan of Karāchi and cover all aspects of city development. GUSTAV RANIS, *Industrial Efficiency and Economic Growth: A Case Study of Karachi* (1961), an analysis of economic development; IMTIAZUDDIN HUSAIN, MOHAMMED AFZAL, and AMJAD ALI BAHADUR RIZVI, *The Social Characteristics of the People of Karachi* (1965), a study of the socioeconomic characteristics of the people; and M. HANIF RAZA, *Karachi: the Show Window of Sind* (1984), remain useful.

Kelvin

William Thomson, who was knighted in 1866 and was raised to the peerage in 1892 (as Baron Kelvin of Largs) in recognition of his work in engineering and physics, was foremost among the small group of British scientists who helped to lay the foundations of modern physics. His contributions to science included a major role in the development of the second law of thermodynamics; the absolute temperature scale (measured in kelvins); the dynamical theory of heat; the mathematical analysis of electricity and magnetism, including the basic ideas for the electromagnetic theory of light; the geophysical determination of the age of the Earth; and fundamental work in hydrodynamics. His theoretical work on submarine telegraphy and his inventions for use on submarine cables aided Britain in capturing a preeminent place in world communication during the 19th century.

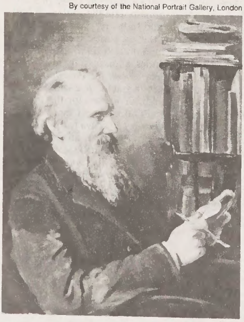
The style and character of Thomson's scientific and engineering work reflected his active personality. While a student at the University of Cambridge, he was awarded silver sculls for winning the university championship in racing single-seater rowing shells. He was an inveterate traveller all of his life, spending much time on the Continent and making several trips to the United States. In later life he commuted between homes in London and Glasgow. Thomson risked his life several times during the laying of the first transatlantic cable.

Thomson's worldview

Thomson's worldview was based in part on the belief that all phenomena that caused force—such as electricity, magnetism, and heat—were the result of invisible material in motion. This belief placed him in the forefront of those scientists who opposed the view that forces were produced by imponderable fluids. By the end of the century, however, Thomson, having persisted in his belief, found himself in opposition to the positivistic outlook that proved to be a prelude to 20th-century quantum mechanics and relativity. Consistency of worldview eventually placed him counter to the mainstream of science.

But Thomson's consistency enabled him to apply a few basic ideas to a number of areas of study. He brought together disparate areas of physics—heat, thermodynamics, mechanics, hydrodynamics, magnetism, and electricity—and thus played a principal role in the great and final synthesis of 19th-century science, which viewed all physical change as energy-related phenomena. Thomson was also the first to suggest that there were mathematical analogies between kinds of energy. His success as a synthesizer of theories about energy places him in the same position in 19th-century physics as Sir Isaac Newton has in 17th-century physics or Albert Einstein in 20th-century physics. All of these great synthesizers prepared the ground for the next grand leap forward in science.

By courtesy of the National Portrait Gallery, London



Kelvin, oil painting by Elizabeth King, 1886–87. In the National Portrait Gallery, London.

Early life. William Thomson was born on June 26, 1824, in Belfast, Ireland, the fourth child in a family of seven. His mother died when he was six years old. His father, James Thomson, who was a textbook writer, taught mathematics, first in Belfast and later as a professor at the University of Glasgow; he taught his sons the most recent mathematics, much of which had not yet become a part of the British university curriculum. An unusually close relationship between a dominant father and a submissive son served to develop William's extraordinary mind.

Influences on his career

William, age 10, and his brother James, age 11, matriculated at the University of Glasgow in 1834. There William was introduced to the advanced and controversial thinking of Jean-Baptiste-Joseph Fourier when one of Thomson's professors loaned him Fourier's pathbreaking book *The Analytical Theory of Heat,* which applied abstract mathematical techniques to the study of heat flow through any solid object. Thomson's first two published articles, which appeared when he was 16 and 17 years old, were a defense of Fourier's work, which was then under attack by British scientists. Thomson was the first to promote the idea that Fourier's mathematics, although applied solely to the flow of heat, could be used in the study of other forms of energy—whether fluids in motion or electricity flowing through a wire.

Influence of Fourier

Thomson won many university awards at Glasgow, and at the age of 15 he won a gold medal for "An Essay on the Figure of the Earth," in which he exhibited exceptional mathematical ability. That essay, highly original in its analysis, served as a source of scientific ideas for Thomson throughout his life. He last consulted the essay just a few months before he died at the age of 83.

Thomson entered Cambridge in 1841 and took his B.A. degree four years later with high honours. In 1845 he was given a copy of George Green's *An Essay on the Application of Mathematical Analysis to the Theories of Electricity and Magnetism.* That work and Fourier's book were the components from which Thomson shaped his worldview and which helped him create his pioneering synthesis of the mathematical relationship between electricity and heat. After finishing at Cambridge, Thomson went to Paris, where he worked in the laboratory of the physicist and chemist Henri-Victor Regnault to gain practical, experimental competence to supplement his theoretical education.

The chair of natural philosophy (later called physics) at the University of Glasgow fell vacant in 1846. Thomson's father then mounted a carefully planned and energetic campaign to have his son named to the position, and at the age of 22 William was unanimously elected to it. Despite blandishments from Cambridge, Thomson remained at Glasgow for the rest of his career. He resigned his university chair in 1899, at the age of 75, after 53 years of a fruitful and happy association with the institution. He was making room, he said, for younger men.

Thomson's scientific work was guided by the conviction that the various theories dealing with matter and energy were converging toward one great, unified theory. He pursued the goal of a unified theory even though he doubted that it was attainable in his lifetime or ever. The basis for Thomson's conviction was the cumulative impression obtained from experiments showing the interrelation of forms of energy. By the middle of the 19th century it had been shown that magnetism and electricity, electromagnetism, and light were related, and Thomson had shown by mathematical analogy that there was a relationship between hydrodynamic phenomena and an electric current flowing through wires. James Prescott Joule also claimed that there was a relationship between mechanical motion and heat, and his idea became the basis for the science of thermodynamics.

Joule's theory of interconvertibility

In 1847 Thomson first heard Joule's theory about the interconvertibility of heat and motion at a meeting of the British Association for the Advancement of Science. Joule's theory went counter to the accepted knowledge of the time, which was that heat was an imponderable substance (caloric) and could not be, as Joule claimed, a form of motion. Thomson was open-minded enough to discuss with Joule the implications of the new theory. At the time, though he could not accept Joule's idea, Thomson was willing to reserve judgment, especially since the relation between heat and mechanical motion fit into his own view of the causes of force. By 1851 Thomson was able to give public recognition to Joule's theory, along with a cautious endorsement in a major mathematical treatise, "On the Dynamical Theory of Heat." Thomson's essay contained his version of the second law of thermodynamics, which was a major step toward the unification of scientific theories.

Thomson's work on electricity and magnetism also began during his student days at Cambridge. When, much later, James Clerk Maxwell decided to undertake research in magnetism and electricity, he read all of Thomson's papers on the subject and adopted Thomson as his mentor. Maxwell—in his attempt to synthesize all that was known about the interrelationship of electricity, magnetism, and light—developed his monumental electromagnetic theory of light, probably the most significant achievement of 19th-century science. This theory had its genesis in Thomson's work, and Maxwell readily acknowledged his debt.

Thomson's contributions to 19th-century science were many. He advanced the ideas of Michael Faraday, Fourier, Joule, and others. Using mathematical analysis, Thomson drew generalizations from experimental results. He formulated the concept that was to be generalized into the dynamic theory of energy. He also collaborated with a number of leading scientists of the time, among them Sir George Gabriel Stokes, Hermann von Helmholtz, Peter Guthrie Tait, and Joule. With these partners, he advanced the frontiers of science in several areas, particularly hydrodynamics. Furthermore, Thomson originated the mathematical analogy between the flow of heat in solid bodies and the flow of electricity in conductors.

The transatlantic cable

Thomson's involvement in a controversy over the feasibility of laying a transatlantic cable changed the course of his professional work. His work on the project began in 1854 when Stokes, a lifelong correspondent on scientific matters, asked for a theoretical explanation of the apparent delay in an electric current passing through a long cable. In his reply, Thomson referred to his early paper "On the Uniform Motion of Heat in Homogeneous Solid Bodies, and its Connexion with the Mathematical Theory of Electricity" (1842). Thomson's idea about the mathematical analogy between heat flow and electric current worked well in his analysis of the problem of sending telegraph messages through the planned 3,000-mile (4,800-kilometre) cable. His equations describing the flow of heat through a solid wire proved applicable to questions about the velocity of a current in a cable.

The publication of Thomson's reply to Stokes prompted a rebuttal by E.O.W. Whitehouse, the Atlantic Telegraph Company's chief electrician. Whitehouse claimed that practical experience refuted Thomson's theoretical findings, and for a time Whitehouse's view prevailed with the directors of the company. Despite their disagreement, Thomson participated, as chief consultant, in the hazardous early cable-laying expeditions. In 1858 Thomson patented his telegraph receiver, called a mirror galvanometer, for use on the Atlantic cable. (The device, along with his later modification called the siphon recorder, came to be used on most of the worldwide network of submarine cables.) Eventually the directors of the Atlantic Telegraph Company fired Whitehouse, adopted Thomson's suggestions for the design of the cable, and decided in favour of the mirror galvanometer. Thomson was knighted in 1866 by Queen Victoria for his work.

Later life. After the successful laying of the transatlantic cable, Thomson became a partner in two engineering consulting firms, which played a major role in the planning and construction of submarine cables during the frenzied era of expansion that resulted in a global network of telegraph communication. Thomson became a wealthy man who could afford a 126-ton yacht and a baronial estate.

Thomson's interests in science included not only electricity, magnetism, thermodynamics, and hydrodynamics but also geophysical questions about tides, the shape of the Earth, atmospheric electricity, thermal studies of the ground, the Earth's rotation, and geomagnetism. He also entered the controversy over Charles Darwin's theory of evolution. Thomson opposed Darwin, remaining "on the side of the angels."

Investigations of the age of the Earth

Thomson challenged the views on geologic and biological change of the early uniformitarians, including Darwin, who claimed that the Earth and its life had evolved over an incalculable number of years, during which the forces of nature always operated as at present. On the basis of thermodynamic theory and Fourier's studies, Thomson estimated in 1862 that more than one million years ago the Sun's heat and the temperature of the Earth must have been considerably greater and that these conditions had produced violent storms and floods and an entirely different type of vegetation. His views, published in 1868, particularly angered Darwin's supporters. Thomas Henry Huxley replied to Thomson in the 1869 Anniversary Address of the President of the Geological Society of London. Thomson's speculations as to the age of the Earth and the Sun were inaccurate, but he did succeed in pressing his contention that biological and geologic theory had to conform to the well-established theories of physics.

In an 1884 series of lectures at Johns Hopkins University on the state of scientific knowledge, Thomson wondered aloud about the failures of the wave theory of light to explain certain phenomena. His interest in the sea, roused aboard his yacht, the *Lalla Rookh,* resulted in a number of patents: a compass that was adopted by the British Admiralty; a form of analog computer for measuring tides in a harbour and for calculating tide tables for any hour, past or future; and sounding equipment. He established a company to manufacture these items and a number of electrical measuring devices. Like his father, he published a textbook, *Treatise on Natural Philosophy* (1867), a work on physics coauthored with Tait that helped shape the thinking of a generation of physicists.

Thomson was said to be entitled to more letters after his name than any man in the Commonwealth. He received honorary degrees from universities throughout the world and was lauded by engineering societies and scientific organizations. Elected a fellow of the Royal Society in 1851, he served as its president from 1890 to 1895. He published more than 600 papers and was granted dozens of patents. He died on Dec. 17, 1907, at Netherhall, his estate near Largs, North Ayrshire, Scotland, and was buried in Westminster Abbey, London.

BIBLIOGRAPHY. HAROLD ISSADORE SHARLIN and TIBY SHARLIN, *Lord Kelvin* (1979), discusses the sources of Thomson's originality. SILVANUS P. THOMPSON, *The Life of Lord Kelvin,* 2nd ed., 2 vol. (1976), includes a complete bibliography of his published works. CROSBIE SMITH and M. NORTON WISE, *Energy and Empire: A Biographical Study of Lord Kelvin* (1989), connects Thomson's scientific and industrial work. DAVID B. WILSON, *Kelvin and Stokes: A Comparative Study in Victorian Physics* (1987), explores the lives of these two friends and scientific collaborators. ROBERT KARGON and PETER ACHINSTEIN (eds.), *Kelvin's Baltimore Lectures and Modern Theoretical Physics* (1987), reproduces Thomson's original wave theory and molecular dynamics lectures in full and includes essays on the larger context of his science. JOE D. BURCHFIELD, *Lord Kelvin and the Age of the Earth* (1975, reprinted 1990), explains Thomson's role in the argument between Darwin and British geologists. PAUL TUNBRIDGE, *Lord Kelvin: His Influence on Electrical Measurements and Units* (1992), focuses on Thomson's efforts to set international standards for electrical measurements.

(H.I.S./Ed.)

Kepler

The Renaissance astronomer and astrologer Johannes Kepler is best known for his discovery of the three principles of planetary motion, by which he clarified the spatial organization of the solar system. Moreover, he founded modern optics by presenting the earliest correct explanation of how human beings see. He was the first to set forth accurately what happens to light after it enters a telescope, and he designed a particular form of that instrument. His ideas provided a transition from the ancient geometrical description of the heavens to modern dynamical astronomy, into which he introduced the concept of physical force.

Early life. On December 27, 1571, in the German town of Weil der Stadt, then a "free city" within the Holy Roman Empire, Johannes was born prematurely, the offspring of an unhappy marriage. His father was a ne'er-do-well mercenary soldier, his mother the quarrelsome daughter of an innkeeper. Small in stature, Johannes never enjoyed robust health, but his superior intelligence was recognized even when he was a young child. Coming from a poor family, he would have received no education had not the dukes of Württemberg adopted the enlightened policy of providing generous scholarships for the bright sons of their impoverished subjects.

With such help Kepler in 1587 was able to attend the University of Tübingen, where he had the good fortune to study astronomy under Michael Mästlin, a professor who may have been unique in his day, for he was convinced that the astronomical system propounded by Nicolaus Copernicus was basically true: the Earth is a planet that rotates daily around its own axis and revolves annually around the Sun. Kepler's youthful acceptance of Copernican astronomy profoundly affected the subsequent course of his life.

Major achievements. After obtaining the B.A. in 1588 and the M.A. in 1591, Kepler planned to become a Lutheran minister. But in 1594, during his last year of training in theology at Tübingen, the teacher of mathematics in the Lutheran high school of Graz, in Austria, having died, Kepler was strongly recommended by the Tübingen faculty to fill the vacancy. Kepler did not finish the theology course at Tübingen but went to Graz the same year. On a summer day in 1595, while he was teaching a class, a spectacular idea flashed through his mind. Ancient Greek geometry had proved that there were five regular solids, or "Platonic bodies": tetrahedron (pyramid), cube, octahedron (formed by eight equilateral triangles), dodecahedron (12 pentagons), and icosahedron (20 equilateral triangles). The ancients knew that these five solids could be enclosed in a sphere, and that there can be no additional regular solids. Sustained by a vision of mathematical harmonies in the skies, a vision he derived from the philosophy of Plato and the mathematics of the Pythagoreans, Kepler tried to relate planetary orbits with geometrical figures.

Kepler, engraving by an unknown artist, *c.* 1730, after a contemporary painting.
Archiv fur Kunst und Geschichte

The notion of cosmic harmony

According to Copernican astronomy there were six planets, whose orbits were regulated by the turning of invisible spheres. But why were there only six planets and not nine or 100? Was the cosmos so constructed that one of the five regular solids intervened between each pair of the unseen spheres, which carried the six Copernican planets? This nest of alternating planets and regular solids constituted the main theme in Kepler's *Prodromus Dissertationum Mathematicarum Continens Mysterium Cosmographicum* ("Cosmographic Mystery"), which he published in 1596 under the auspices of the Tübingen faculty. The Platonic and Pythagorean components in Kepler's conception of celestial harmony, however mystical in origin, helped to lead him to the three principles of planetary motion now known by his name.

Kepler sent copies of his first major work to a number of scientists, including Tycho Brahe, who was soon to become the imperial mathematician of the Holy Roman Empire. Although Brahe did not agree with the underlying Copernican foundation of Kepler's *Mysterium Cosmographicum,* he was so impressed by the author's knowledge of astronomy and skill in mathematics that in 1600 he invited him to join his research staff in the observatory at Benatek (now Benátky nad Jazerou), outside Prague. When Brahe died the next year, Kepler was promptly appointed his successor as imperial mathematician. His first publication at Prague, *De Fundamentis Astrologiae Certioribus* (1601; "The More Reliable Bases of Astrology"), rejected the superstitious view that the stars guide the lives of human beings. Nonetheless, his deep feeling for the harmony of the universe included a belief in the harmony between the universe and the individual, and his skill in astrological prediction was much in demand.

While Kepler was watching a rare conjunction of Mars, Jupiter, and Saturn in October 1604, a supernova appeared that remained visible for 17 months. This event was evidence that the realm of the fixed stars, considered since ancient times as pure and changeless, could indeed experience change. He published the results of his observations in 1606 as *De Stella Nova in Pede Serpentarii* ("The New Star in the Foot of the Serpent Bearer").

Contributions to optics

Kepler now had access to Brahe's incomparable collection of astronomical observations, the result of decades of unremitting and painstaking toil by the greatest naked-eye observer of the heavens and the leader of a highly qualified team of astronomers. As a member of the team, Kepler had been assigned to investigate the planet Mars. But, before he could use the raw observations, Kepler felt that he had to solve the problem of atmospheric refraction: how is a ray of light, coming from a distant heavenly body located in the less dense regions of outer space, deflected when it enters the denser atmosphere surrounding the Earth?

Kepler incorporated his results in a book that he modestly entitled *Ad Vitellionem Paralipomena, Quibus Astronomiae Pars Optica Traditur,* (1604; "Supplement to Witelo, Expounding the Optical Part of Astronomy"); Witelo (Latin Vitellio) had written the most important medieval treatise on optics. But Kepler did much more than add to his work. He made an analysis of the process of vision that provided the foundation for all of the advances in the understanding of the structure and function of the human eye. Kepler wrote that every point on a luminous body in the field of vision emits rays of light in all directions, but that only those rays can enter the eye that impinge on the pupil, which functions as a diaphragm. He stated that the rays emanating from a single luminous point form a cone, the circular base of which is in the pupil. All of the rays are then refracted within the normal eye to meet again at a single point on the retina, identified by Kepler as the sensitive receptor of the eye. If the eye is not normal, the second short interior cone comes to a point not on the retina but in front of it or behind it, causing blurred vision. For more than three centuries eyeglasses had helped older persons to see better. But nobody before Kepler was able to explain how these little pieces of curved glass had worked.

After the invention of the telescope had been reported to Galileo, who promptly proceeded to make his astounding discoveries, Kepler applied the same ideas concerning optics to the explanation of how the telescope works. Although Galileo's findings were received in general with skepticism and ridicule, Kepler acknowledged the Italian's accomplishments in his *Dissertatio cum Nuncio Sidereo Nuper ad Mortales Misso a Galilaeo Galilaeo* in 1610.

Galileo did not return the compliment. He chose to ignore the epoch-making results Kepler had published the preceding year. In his *Astronomia Nova* ("New Astronomy") of 1609, Kepler had demonstrated that the orbit of the planet Mars is an ellipse. Although it had been believed since antiquity that the planets, being heavenly bodies, were perfect and therefore could move only in perfect circles or combinations of circles, Copernicus had correctly classified the Earth as one of the planets; and it was fully accepted that the Earth was far removed from perfection. Kepler extended Copernicus' reasoning to the other planets and was the first to declare that the other planets resemble the Earth in being material bodies. That a material body, being imperfect, need not travel in a perfectly circular orbit was a conclusion made by Kepler after he tried unsuccessfully to fit the orbit of Mars to Brahe's observations in every possible combination of circles his ingenuity could devise. Because none of them worked, he tried noncircular paths until he found the true solution: Mars revolves in an elliptical orbit with the Sun occupying one of its two focuses.

Kepler's three principles of planetary motion

The pre-Keplerian dogma that permitted only circular paths entailed the concept of uniform motion—*i.e.*, the moving body or point must traverse equal arcs in equal intervals of time. Such a conception of uniform motion as measured along an arc was, of course, incompatible with an elliptical orbit. But Kepler found an alternative form of uniformity. This new uniformity equated equal areas with equal times. With the Sun remaining stationary in one focus of the ellipse, the planet, while revolving along the periphery of its elliptical orbit, would sweep out, in equal intervals of time, equal areas of the ellipse, not equal arcs along the periphery of the ellipse.

In 1619, 10 years after Kepler published these first two principles of planetary motion (the elliptical orbit and equality of areas), he published the *Harmonice Mundi* (*Harmonics of the World*), in which he expounded his third principle, which related a planet's mean distance from the Sun to the time it takes to complete its elliptical orbit around the Sun. The cube of the distance proved to be in a constant ratio to the square of the time required for all the planets to complete such an orbit. The enunciation of this rule (which is sometimes called the $^3/_2$ ratio) completed Kepler's contribution to the understanding of planetary motion and helped to prepare the way for Sir Isaac Newton's exposition of universal gravitation, which affects all of the material bodies in the physical universe.

Later life. Meanwhile, Kepler's patron, the Holy Roman Emperor, had been compelled by his brother to abdicate, and Kepler himself had found it desirable to leave Prague, then the capital of the empire. Although he was reappointed imperial mathematician by the new emperor, Kepler moved to Linz, in Austria. His first wife had died in Prague; Kepler remarried in 1613. Once, when buying supplies for his new home, Kepler became unhappy about the rough-and-ready methods used by the merchants to estimate the liquid contents of a wine barrel. Because the curved containers they used were of various shapes, Kepler sought a mathematical method for determining their volumes. Following the model established by Archimedes, the most talented mathematician of antiquity, Kepler, in his volumetric researches, investigated the properties of nearly 100 solids of revolution—made by rotating a two-dimensional surface on one of its axes—that had not been considered by Archimedes. Starting with an ordinary wine barrel, Kepler enormously extended the range of Archimedes' results. He did so by refusing to confine himself, as Archimedes had done, to cases in which a surface is generated by a conic section—a curve formed by the intersection of a plane and a cone—rotating about its principal axis. Kepler's additional solids are generated by rotation about lines in the plane of the conic section other than its principal axis.

While he was in Linz, Kepler published his *Epitome Astronomiae Copernicanae* (1618–21; *Epitome of Copernican Astronomy*). He modelled this title after the highly successful introduction to astronomy that had been published by his former Tübingen professor in a number of editions. But, whereas Mästlin had deemed it prudent pedagogical practice to keep Copernicanism out of an elementary textbook, which he therefore entitled simply "Epitome of Astronomy," Kepler emphasized his open espousal of the new cosmology by inserting the provocative label "Copernican."

Witchcraft trial of his mother

In Linz in 1620, Kepler heard that his mother had been indicted on the charge of being a witch. Such a defendant was often subjected to torture and, if convicted, was usually burned at the stake. If his mother had suffered this fate, Kepler's own status as imperial mathematician of the Holy Roman Empire and mathematician of Upper Austria might have been irreparably impaired. He rushed to her defense, therefore, not only out of filial devotion but also out of prudent self-interest. Only his skillful intervention saved her from torture and a fiery death.

Kepler had planned to publish his *Tabulae Rudolphinae* (*Rudolphine Tables*), named in honour of his first imperial patron, Rudolph II, in Linz. But this work, the final outcome of long years of unceasing reflection and tireless calculation, could not be printed there because of a rebellion by the peasants, who were infuriated by a combination of being forced to return to Catholicism and to pay heavy taxes. Kepler had to find another home and a new patron.

Albrecht von Wallenstein, duke of Friedland and Żagań, a successful soldier of fortune who had put his private army at the disposal of the empire in the Thirty Years' War, accepted the responsibility of satisfying Kepler's financial needs. The astronomer moved to Żagań in Silesia, where he was able to establish his own printing press. The *Rudolphine Tables* were printed at Ulm, Germany, in 1627, before Kepler went to Żagań in 1628. But Wallenstein turned out to be someone on whom Kepler could not rely.

Leaving his family behind in Żagań, Kepler went west to collect the interest due on two promissory notes he held in exchange for money he had deposited in Austria. On his way he stopped at Regensburg, where the Imperial Diet was in session. He fell acutely ill and died on November 15, 1630. The tremendous upheavals suffered by Germany in the Thirty Years' War later obliterated his grave.

BIBLIOGRAPHY. Kepler's complete works are available in *Gesammelte Werke*, ed. by WALTHER VON DYCK and MAX CASPAR, 19 vol. (1937–81), a magnificent edition, and *Opera Omnia*, ed. by CH. FRISCH, 8 vol. (1858–71), still a useful collection. Selections in English include *Epitome of Copernican Astronomy: IV and V*, and *The Harmonies of the World: V*, both in MORTIMER J. ADLER (ed.), *Great Books of the Western World*, 2nd ed., vol. 15 (1990), *Johannes Kepler: Life and Letters*, ed. and trans. by CAROLA BAUMGARDT (1951), a selection of his letters with an introduction by Albert Einstein on his scientific achievements, *The Six-Cornered Snowflake*, ed. and trans. by COLIN HARDIE (1966), including parallel Latin and English texts and also containing numerous annotations and photographs, *The Secret of the Universe*, trans. by A.M. DUNCAN (1981), also including parallel Latin and English texts, and *New Astronomy*, trans. by WILLIAM H. DONAHUE (1992). MAX CASPAR (ed.), *Bibliographia Kepleriana*, 2nd ed. by MARTHA LIST (1968), is a complete enumeration of the publications by and about Kepler. ARTHUR BEER and PETER BEER (eds.), *Kepler: Four Hundred Years: Proceedings of Conferences Held in Honor of Johannes Kepler* (1975), includes List's supplement to the *Bibliographia* covering the years 1967–75.

Biographies are MAX CASPAR, *Kepler* (1959, reissued 1993; originally published in German, 3rd ed., 1958); ARTHUR KOESTLER, *The Watershed: A Biography of Johannes Kepler* (1960, reprinted 1984); and A. ARMITAGE, *John Kepler* (1966), contrasting Kepler's thought with medieval superstitions.

Kepler's science is investigated by ROBERT SMALL, *An Account of the Astronomical Discoveries of Kepler* (1804, reprinted 1963); EDWARD ROSEN (ed. and trans.), *Kepler's Conversation with Galileo's Sidereal Messenger* (1965), on Kepler's enthusiastic reaction to Galileo's spectacular discoveries with the telescope, and *Somnium* (1967), on Kepler's imaginary space flight to the Moon; and JUDITH VERONICA FIELD, *Kepler's Geometrical Cosmology* (1988), which analyzes the influence of Kepler's astrological, philosophical, and religious beliefs on his work in the field of cosmology. Specialized studies on Kepler include EDWARD ROSEN, "In Defense of Kepler," in ARCHIBALD R. LEWIS (ed.), *Aspects of the Renaissance: A Symposium* (1967), pp. 141–158, "Galileo and Kepler: Their First Two Contacts," *Isis*, 57:262–264 (1966), "Kepler's *Harmonics* and His Concept of Inertia," *American Journal of Physics*, 34:610–613 (1966), and "Kepler and Witchcraft Trials," *The Historian*, 28:447–450 (1966). ALEXANDRE KOYRÉ, *The Astronomical Revolution: Copernicus, Kepler, Borelli* (1973, reissued 1992; originally published in French, 1961); and OWEN GINGERICH, *The Eye of Heaven: Ptolemy, Copernicus, Kepler* (1993), study of the evolution of astronomy. NICHOLAS JARDINE, *The Birth of History and Philosophy of Science: Kepler's A Defence of Tycho Against Ursus, With Essays on its Provenance and Significance* (1984); and EDWARD ROSEN, *Three Imperial Mathematicians: Kepler Trapped Between Tycho Brahe and Ursus* (1986), recount a controversial episode in the history of early modern science.

(Ed.R./Ed.)

Kiev

The capital and largest city of Ukraine, a port on the Dnieper (Dnipro) River, and a large railroad junction, Kiev (Kyyiv) is a city with an ancient and proud history. As the centre of Kievan Rus, the first eastern Slavic state, 1,000 years ago it acquired the title "Mother of Rus Cities." It was severely damaged during World War II, but by the mid-1950s it was fully restored, and by the 1970s it had become a thriving, modern city with a well-developed economic and cultural life. The emergence of an independent Ukraine in the early 1990s renewed Kiev's status as a major European capital.

The article is divided into the following sections:

Physical and human geography

THE LANDSCAPE

Site. The city stands on the Dnieper River just below its confluence with the Desna and 591 miles (951 kilometres) from its mouth in the Black Sea. The original location was on the high and steep right bank, which rises above the river in an imposing line of bluffs culminating in Batyyeva Hill 330 feet (100 metres) above mean river level. This precipitous and wooded bank, topped by the golden domes and spires of churches and bell towers and by modern high-rise apartment buildings, makes the city an attractive and impressive sight from across the Dnieper. Since World War II, Kiev has extended onto the wide, low, and flat floodplain on the left bank.

Climate. Kiev has a moderately continental climate. The average January temperature is 21° F (−6° C), and winter days with temperatures above freezing are not uncommon; in cold spells with a northerly or northeasterly airstream, temperatures may drop sharply, and an absolute minimum of −27° F (−33° C) has been recorded. Snow cover lies usually from mid-November to the end of March; the frost-free period lasts 180 days on average but in some years surpasses 200 days. Summers are warm, with a July average of 68° F (20° C) and a recorded maximum of 102° F (39° C). The mean annual precipitation is 25 inches (635 millimetres), with maximum rainfall in June and July.

Layout. The city limits enclose an area of 300 square miles (780 square kilometres) on both banks of the Dnieper. It is divided into 12 administrative wards. The focus of Kiev is the area of the ancient Upper Town, crowning the high bluffs of the Dnieper. Although largely of postwar construction, this central area retains its old street pattern, and most of the surviving historical and architectural monuments are located there. First among these is the Cathedral of St. Sophia, now a museum. It was founded in the 11th century and remains, despite certain Baroque modifications in the 18th century, one of the finest and most beautiful examples of early Russo-Byzantine ecclesiastical architecture. It has a nave and four aisles and is crowned by five domes. The interior is magnificently decorated with frescoes and mosaics; it contains the tomb of Yaroslav, during whose reign the cathedral was built.

The Upper Town

Close by is the Baroque Church of St. Andrew, designed by Bartolomeo Rastrelli and built in the mid-18th century; its site on the crest of the steep slope to the river makes it a striking landmark. Other historical relics in the central area include the ruins of the Golden Gate, also built in the 11th century in the reign of Yaroslav; the Zaborovskyy Gate, built in 1746–48; and the remains of the Desyatynna church, or Church of the Tithes, built in 989–996 by St. Volodymyr (Vladimir).

Within and immediately adjacent to the area of the former Old Town are many of the city's museums, theatres, and public buildings as well as the principal shops, including the central department store and the covered market. The axis of the centre is the street known as Khreshchatyk, which runs along the bottom of a small valley the sides of which have in part been landscaped with terraced gardens interspersed with tall, modern office and apartment buildings. The greenery of the gardens, the trees lining the street, the squares that it intersects—all combine with the variegated colours of brick, red and gray granites, and decorative ceramic tiles to give Khreshchatyk an attractive and colourful aspect much admired by Kiev's inhabitants. Among important buildings on the street is that of the city council, where the 800 elected deputies hold their meetings.

Intersecting Khreshchatyk at right angles is the wide, poplar-lined Boulevard of Taras Shevchenko, on which stands the university with its eye-catching red-washed walls. There, too, is the Cathedral of St. Volodymyr (still in use as a church), built in 1850–96 in Byzantine style and containing impressive paintings by Viktor Vasnetsov and other Russian artists. Notable among the many statues in central Kiev are those that commemorate the Cossack leader Bohdan Khmelnytsky and the Ukrainian poet Taras Shevchenko.

North of the old centre is the former trading and Jewish quarter, Podil, with a rectangular pattern of streets and the old merchants' trading exchange, the House of Contracts, built in 1817. Also north of the old centre is the river port. South of the centre is the Pecherskyy district, along the top of the riverbank. This district contains many of the principal buildings of the Ukrainian government, including the glass-domed palace, built in 1936–39, that houses the Supreme Council and the 10-story block that houses the Cabinet of Ministers. Nearby is the attractive Mariinskyy Palace, built in 1752–55 for the *tsaritsa* Elizabeth, reconstructed in 1870, and now used for government receptions.

The Pecherskyy district

At the southern end of this district is the Kyevo-Pecherska Lavra (Monastery of the Caves), founded in the early 11th century, one of the most famous and important monasteries in the history of the region. It was at the *lavra* that the monk Nestor wrote the earliest surviving chronicle of the East Slavic state of Rus. Although the Cathedral of the Assumption (inside the walls of the monastery) was blown up in 1941, Trinity Church, of the same period, survives. Also within the walls are the 17th-century Church of All Saints and an impressive 18th-century bell tower rising 315 feet. A major feature of the monastery is the system of catacombs beneath it in which the mummified bodies of early monks and saints, including Nestor, are entombed. Although it is a museum open to the public, the Kyevo-Pecherska Lavra is still in use as a monastery. South from the *lavra* is yet another monastery, the Vydubytskyy, dating from the 11th century; it, too, was severely damaged in World War II.

All along the steep riverbank, fronting the Upper Town and Pecherskyy district, an attractively landscaped park has been laid out overlooking the Dnieper. With the views it affords, the park forms one of the most striking features of the city. It contains an open-air theatre, sports stadium, and restaurant, and a funicular railway climbs the 300-foot slope. Also within the park are many memorials. Dominating the northern end is the statue of Prince Volodymyr, who brought Christianity to Rus. The statue marks the place where in 988 the people of Kiev were baptized en masse. The southern end, called the Park of Glory, has an 85-foot granite obelisk rising above the grave of the Unknown Soldier and a memorial garden. Also located in the park are the grave of General Nikolay Vatutin, commander of the Soviet forces that liberated Kiev in 1943, and a rotunda marking the supposed grave of the early Varangian chief Askold.

Suburbs

Around these central districts of Kiev stretch extensive suburbs of factories and residential neighbourhoods. As low priority was given to housing during the Stalin period, the greater part of these suburbs was built after his death. The neighbourhood units, known as microregions, consist of groupings of apartment buildings housing 2,500 to 5,000 people, together with basic services, local shops, a health centre, cinema, and primary school. Since the late 1960s the apartment buildings have usually been of 12 to 20 stories and of prefabricated construction. Most apartments have only two or three rooms, and population densities are therefore high, in the new residential developments as much as in the older central areas. The growing ownership of private cars poses problems in the provision of garage space in these new districts. A feature of development since World War II has been the rapid spread of the city on the low left bank of the Dnieper, previously almost devoid of settlement. The left bank is linked to the main part of Kiev by a railway bridge and by the imposing Ye.O. Paton road bridge, which is 4,920 feet long and named for its designer.

Between the neighbourhood units are substantial areas of parks and green space. These include the large botanical gardens of the Academy of Sciences of Ukraine, the smaller university botanical gardens (established in the mid-19th century), and in the southwestern suburbs the extensive permanent exhibition of the Ukrainian economy. On the city outskirts are several areas of forest, which are much used for recreation. In the south is the Holosiivskyy Forest Park, dominated by deciduous trees, and to the north are nearly 10,000 acres (4,450 hectares) of the Pushcha-Vodytsya Forest Park, mainly covered by coniferous species. A number of factors combine to make Kiev an attractive urban centre—the site, with its sharply contrasted relief and wide views across the Dnieper, the abundance of greenery in and around the city, and the many buildings of historic interest and beauty.

THE ECONOMY

Industry. Kiev, as the capital of Ukraine, has major administrative functions, with considerable employment in the offices of ministries responsible for the economy. The city is also an important industrial centre, possessing a wide range of manufactures. Factories are found in all quarters of the city, with major concentrations to the west of the city centre and on the left bank of the Dnieper.

Engineering industries, based on metal from the iron and steel plants of the Dnieper Bend region and the Donets Basin (Donbas) coalfield, take pride of place and include the production of complex machinery and precision tools and instruments. Plants in Kiev make equipment for chemical works, such as conveyor lines for vulcanized rubber, linoleum, and fertilizer factories, and also produce metal-cutting machines. Other engineering products are aircraft, hydraulic elevators, electrical instruments, armatures, river- and seacraft, motorcycles, and cinematography apparatus.

Chemical industry

Another important sector is the chemical industry, making resin products, fertilizers, plastics, and chemical fibres, the last at the Darnytsya viscose rayon plant on the left bank. Lumber milling and the making of bricks and reinforced concrete items also are well developed. Consumer goods manufactured include cameras, thermos flasks, knitwear, footwear, a range of foodstuffs, and watches. Kiev is also a large publishing centre.

Power for the many enterprises is supplied by natural gas, piped from Dashava in western Ukraine, and by electricity from the Kiev hydroelectric station on the Dnieper. This station, completed in 1968, is at Vyshhorod, just upstream of the city. Twenty-five miles southeast of Kiev is the still more powerful Trypillya thermal electric station.

Transportation. Transportation for the industries and

Anatoly Poddubny—Tass/Sovfoto

The Bohdan Khmelnytsky Square in Kiev. In the foreground is the Cathedral of St. Sophia, now a museum.

for the city as a whole is provided by a good communications network. Trunk railways and all-weather roads link Kiev to Moscow, to Kharkiv and the Donets Basin, to southern Ukraine and the port of Odessa, and to western Ukraine and Poland. The navigability of the Dnieper has been improved by a series of barrages and reservoirs. Boryspil airport operates direct flights to many Ukrainian towns and international service to major cities throughout Europe, Asia, and North America. Within Kiev itself there is efficient subway and rail, bus, streetcar, and trolleybus service.

SOCIAL AND CULTURAL LIFE

Kiev's ancient tradition as a cultural centre is still vigorously alive. The Kiev T.H. Shevchenko State University heads an array of 20 institutions of higher education, notable among which are the Polytechnic (founded in 1898), the Agricultural Academy, and the medical, art, and architectural institutes.

There is a large number of general secondary schools, evening schools for adults, and specialist technical schools. A range of research establishments is headed by an Academy of Sciences, which also maintains the largest of the city's many libraries. Kiev is noted for medical and cybernetic research. The emphasis on applied research is illustrated by the academy's Ye.O. Paton Institute of Electrical Welding.

There are several theatres, notably the Shevchenko Theatre of Opera and Ballet. Plays are presented at the Lesya Ukrainka and Ivan Franko theatres, which specialize in Russian and Ukrainian drama, respectively; drama is also frequently staged in the 4,000-seat auditorium of the Palace of Culture and in the Palace of Sport, which can seat 12,000 people. In addition there are youth, open-air, and musical comedy theatres. Kiev has a circus and more than 130 cinemas; films are made in a studio in the city. Concerts are regularly given at the Tchaikovsky Conservatory. The most important of the city's many museums are the Kiev State Historical Museum, the Kiev State Museum of Russian Art, and the Kiev State Museum of Ukrainian Art.

Sports

Kiev has good facilities for sports; the largest of its 15 stadiums, the Central Stadium, can accommodate 100,000 people. Aquatic sports take place on the reservoir of the Kiev dam at Vyshhorod and also on Trukhaniv Island in the Dnieper opposite the city centre, where there is a fine beach and water sports centre. The city is well provided with health facilities, including general and specialized hospitals and local polyclinics, the latter serving residential neighbourhoods. Since the majority of women are employed, a number of nursery schools and crèches care for children below school age. Around the outskirts of Kiev are health resorts, sanatoriums, and children's holiday camps.

History

THE EARLY PERIOD

Origins and foundation. Kiev has a long, rich, and often stormy history. Its beginnings are lost in antiquity. Archaeological findings of stone and bone implements, the remains of primitive dwellings built of wood and skins, and large accumulations of mammoths' bones indicate that the first settlements in the vicinity date from the Upper Paleolithic Period (some 15,000 to 40,000 years ago). As early as 3000 BC in the Neolithic Period and subsequently at the time of the Cucuteni-Trypillya culture at the end of the Neolithic, tribes engaged in agriculture and animal husbandry lived on the site of modern Kiev. Excavations continue to uncover many artifacts from settlements dating from the Copper, Bronze, and Iron ages. The tribes of the area traded with the nomadic peoples of the steppes to the south, Scythians, Sarmatians, and later Khazars, and also with the ancient Greek colonies that were located on the Black Sea coast.

According to the 12th-century chronicle *Povest vremennykh let* ("Tales of Bygone Years," also known as the Russian Primary Chronicle), Kiev was founded by three brothers, Kiy, Shchek, and Khoriv, leaders of the Polyane tribe of the East Slavs. Each established his own settlement on a hill, and these became the town of Kiev, named for the eldest brother, Kiy; a small stream nearby was named for their sister Lybed. Although the chronicle account is legendary, there are contemporary references to Kiev in the writings of Byzantine, German, and Arab historians and geographers. Archaeological evidence suggests that Kiev was founded in the 6th or 7th century AD.

The first Rus capital. Less legendary is the chronicle account of the Varangians, who seized Kiev in the mid-9th century. As in Novgorod to the north, a Slavo-Varangian ruling elite developed. Kiev, with its good defensive site on the high river bluffs and as the centre of a rich agricultural area and a group of early Slavic towns, began to gain importance. About 882 Oleg (Oleh), the ruler of Novgorod, captured Kiev and made it his capital, the centre of the first East Slavic state, Kievan Rus. The town flourished, chiefly through trade along the Dnieper going south to Byzantium and north over portages to the rivers flowing to the Baltic, the so-called "road from the Varangians to the Greeks," or "water road." Trade also went to the Caspian Sea and Central Asia.

Christianization

In 988 the introduction of Christianity to Kiev enhanced its significance as the spiritual centre of Rus. By the 12th century, according to the chronicles, the city's wealth and religious importance was attested to by its more than 400 churches. The Cathedral of St. Sophia, parts of the Kyevo-Pecherska Lavra, and the ruins of the Golden Gate remain today as witnesses to Kiev at the height of its splendour. The town was famed for its art, the mosaics and frescoes of its churches, its craftsmanship in silver, and the quality of many of its manufactures. One of Europe's major cities, Kiev established diplomatic relations with Byzantium, England, France, Sweden, and other countries. Travelers wrote of its population as numbering tens of thousands.

Throughout the period of Kievan Rus, however, the city was engaged in a succession of wars against the nomadic warrior peoples who inhabited the steppes to the south, in turn the Khazars, Pechenegs, and Polovtsy (Kipchaks). These conflicts weakened the city, but even greater harm was done by the endless, complex internecine struggles of the princedoms into which Rus was divided. In 1169 Prince Andrew Bogolyubsky of Rostov-Suzdal captured and sacked Kiev. Thus by the late 12th century the power of the city had declined, and in the following century it was unable to resist the rising and formidable power of the Mongols. In 1238 a Mongol army under Batu, grandson of Genghis Khan, invaded Rus and, having sacked the towns of central Rus, in 1240 besieged and stormed Kiev. Much of the city was destroyed and most of its population killed. The Franciscan friar and traveler Giovanni da Pian del Carpini six years later reported only 200 houses surviving in Kiev.

Kiev under Lithuania and Poland. In the 14th century what was left of Kiev and its surrounding area came under the control of the powerful and expanding grand duchy of Lithuania, which captured it in 1362. For a long time thereafter Kiev had little function except as a fortress and minor market on the vaguely defined frontier between Lithuania and the steppe Tatars, based in the Crimea. It frequently came under attack from the Tatars; in 1482 the Crimean khan, Mengli Giray, took and sacked the town. Almost the only survival of Kiev's former greatness was its role as the seat of an Orthodox metropolitan. A step forward came in 1516, when the grand duke Sigismund I granted Kiev a charter of autonomy, thereby much stimulating trade.

In 1569 the Union of Lublin between Lithuania and Poland gave Kiev and the Ukrainian lands to Poland. Kiev became one of the centres of Orthodox opposition to the expansion of Polish Roman Catholic influence, spearheaded by vigorous proselytization by the Jesuits. In the 17th century a religious Ukrainian brotherhood was established in Kiev, as in other Ukrainian towns, to further this opposition and encourage Ukrainian nationalism. Peter Mogila (Petro Mohyla), a major theologian and metropolitan of Kiev from 1633 to 1646, founded there the Collegium (later the Academy of Kiev), which became a major focus of the struggle with Roman Catholicism.

In the 17th century there was also increasing unrest among the Zaporozhian Cossacks of the Dnieper downstream of Kiev and an ever-growing struggle between them and the Polish crown. This eventually culminated in the revolt of Bohdan Khmelnytsky, who, assisted by the Crimean Tatars, entered Kiev with his insurgent Cossacks in 1648. He came under heavy pressure from the Polish forces, and in 1654 Khmelnytsky and the Cossacks signed the Pereyaslav Agreement, in essence submitting Ukraine to Moscow; this was followed by a prolonged and confused period of strife and destruction leading in 1667 to the Treaty of Andrusovo, by which Kiev and the Dnieper left-bank part of Ukraine became an autonomous Cossack state under the suzerainty and protection of Moscow. Thereafter further struggle ensued against the Turks, with the Cossacks constantly changing sides and engaging in internecine disputes. In 1686 Kiev was finally yielded to Muscovy by Poland and stood as the sole Muscovite outpost on the right bank of the Dnieper.

The Cossack state

EVOLUTION OF THE MODERN CITY

Kiev under the tsars. In 1793 the Second Partition of Poland, under Catherine the Great, brought right-bank Ukraine into the Russian Empire, and Kiev, assisted by the abolition in 1754 of the tariff barriers between Russia and the Ukrainian lands, began to grow in commercial importance. Catherine's reign was marked by the abolition of the old administrative system and of the post of Cossack hetman and the division of Ukraine into new administrative provinces, for one of which Kiev became the centre. Subsequently it became the centre of a governor-generalship covering three provinces.

In the first half of the 19th century, Kiev developed as a major focus of Ukrainian nationalism, although severe persecution from the tsarist government forced the movement to shift the brunt of its activities to Lviv in the Austrian-ruled Ukrainian regions. In Kiev, as in Russian cities, there was clandestine revolutionary activity (beginning with the Decembrists in the early 19th century) that culminated in a series of strikes and demonstrations leading to the Russian Revolution of 1905. An important role in this revolutionary movement was taken by students of the University of Kiev (now Kiev T.H. Shevchenko State University), which had been established in 1834.

During the 19th century the expanding economic importance of Ukraine, and especially the growing export of grain, brought further commercial development to Kiev. Modern factory industry appeared; to the Arsenal, which had been set up as early as the 18th century, were added lumber milling and the building of rivercraft. The town developed significant industries processing agricultural products—leather, tobacco, distilling, brewing, and textiles. In the late 1860s Kiev was connected by rail to both Moscow and the Black Sea port of Odessa, further enhancing its role as a centre of industry, commerce, and administration. By the outbreak of World War I, the city had a population of some 350,000.

The revolutionary period. With the outbreak of the Russian Revolution of 1917, a revolutionary soviet, the Central Rada (*rada,* "council"), was elected by the city workers, consisting primarily of Menshevik and Socialist Revolutionary members, with strong support from Ukrainian nationalist groups. In January 1918 the Rada proclaimed an independent Ukrainian state with Kiev as its capital. Minor uprisings by Bolshevik workers, who were mostly concentrated in the Arsenal works, were suppressed, but Red Army troops came to their aid and on Feb. 8, 1918, entered Kiev.

By the Treaty of Brest-Litovsk (March 3) between the Bolshevik government and the Germans, however, the new Soviet government recognized the independence of Ukraine, which was promptly occupied by German troops. A puppet Ukrainian government was set up in Kiev by the Germans, but it collapsed with the German surrender to the Allies in November 1918 and the subsequent withdrawal of German troops. Once more an independent Ukraine was declared in Kiev, under the leadership of Simon Petlyura, but its brief and stormy history was a series of struggles between Ukrainian nationalist, White, and Red forces. In November 1919 Kiev was briefly taken by the White armies under General A.I. Denikin before being finally occupied by the Red Army. Peace was still denied the city, with the outbreak of the Russo-Polish War. In May 1920 the Poles captured Kiev but were driven out in a counterattack.

The Soviet period. Kiev's role as the centre for Ukrainian nationalists caused the Soviet government to transfer the capital of the new Ukrainian Soviet Socialist Republic to Kharkiv, and it was not until 1934 that Kiev resumed its capital status. Meanwhile, restoration of the city's shattered economy was undertaken. During the first five-year plans, between 1928 and 1940, new machine tool, electrical, and chemical industries were established. By 1939 the population had reached 846,724. The German invasion in 1941 again brought severe suffering and destruction to the city. After a fierce 80-day battle, German forces entered it on Sept. 19, 1941. More than 30,000 Jews, Soviet prisoners of war, and partisans who had remained in the city were massacred within days in a nearby ravine known as Baby Yar; tens of thousands more were killed there over the next two years. Many of Kiev's other inhabitants were deported for forced labour and to concentration camps, including almost all the large prewar Jewish population. In 1943 the advancing Soviet troops forded the Dnieper and, after bitter fighting, liberated Kiev on November 6. The city itself had suffered great destruction, including more than 40 percent of its buildings and some 800 of its industrial enterprises. For its role in the war, Kiev was later honoured by the Soviet government with the Order of Lenin, the title of Hero-City, and the Gold Star medal. In the first postwar five-year plan, rapid reconstruction was undertaken.

Kiev continued to grow and to strengthen its industrial base during the mid- and late 20th century. Whereas during the Soviet period Kiev as an international political entity fell largely under the shadow of Moscow, the establishment of an independent Ukraine in the early 1990s returned Kiev to the world political stage.

Post-Soviet Kiev

BIBLIOGRAPHY. O.K. KASYMENKO (ed.), *Istoriia Kieva,* 2 vol. (1963–64), is a thorough historical study based on archaeological evidence as well as documentary sources. MICHAEL F. HAMM, *Kiev: A Portrait, 1800–1917* (1993), includes an extensive bibliography. A.V. KUDRYTS'KYĬ (ed.), *Kiev: entsiklopedicheskiĭ spravochnik,* 3rd ed., enlarged (1986), is an encyclopaedic handbook. LEONID DAEN, PAVEL POZNYAK, and MARK CHERP, *Kiev: Travel Guide,* trans. from Ukrainian (1971); and H. LEVITSKY, *Kiev: A Short Guide,* trans. from Russian (1980), are useful descriptive works. (R.A.F./Ed.)

Kinshasa

Kinshasa, the capital of the Democratic Republic of the Congo, lies about 320 miles (515 kilometres) from the Atlantic Ocean on the south bank of the Congo River. One of the largest cities of sub-Saharan Africa, it is a special political unit equivalent to a Congolese region, with its own governor. The city's inhabitants are popularly known as Kinois.

Kinshasa is distinctive not only as the capital of Congo but also as the centre of the dynamic and contradictory influences that have shaped the country's character in modern Africa. The only city not clearly identified with any particular region of the country, it was (until 1997) the seat of the long-lasting Zairean military government based, on the one hand, on the strength of the armed forces and, on the other, on a technique of political and social compromise that until its later years gained the rather grudging collaboration of most of the citizens. Caught between spectacular wealth and massive poverty, most Kinois must spend a considerable amount of their time scrambling for necessities that are in erratic supply. Nevertheless, they have found the means to make Kinshasa a source of distinctive influence in intellectual and popular culture that has been felt throughout Africa.

This article is divided into the following sections:

Physical and human geography

THE LANDSCAPE

The city site and climate. Kinshasa spreads out southward from the shoreline of the Congo River at Malebo (Stanley) Pool, a widening of the river. The plain on which the city lies varies mostly between 918 and 1,148 feet (280 and 350 metres) above sea level and is partly encircled by higher ground. The most heavily inhabited area of Kinshasa covers 58 square miles (about 150 square kilometres). The total area subject to city government, much of it sparsely populated, is 3,847 square miles.

The climate is hot year-round, with a dry season from May to September and a rainy season from October to May. The mean annual rainfall is slightly more than 60 inches (1,524 millimetres). Violent rainstorms occur frequently but seldom last more than a few hours. The hottest month is April, with mean daily maximum and minimum temperatures of 89° F (32° C) and 71° F (22° C), respectively. The corresponding figures for July, the coldest month, are 81° F (27° C) and 64° F (18° C). The higher suburbs are somewhat cooler than the central city. The surrounding countryside is heavily farmed savanna and gallery forest; the chief crops are cassava, sugarcane, oil palms, plantains, corn (maize), peanuts (groundnuts), and beans.

The city layout. The built-up area of Kinshasa is divided into industrial, residential, and commercial zones. Along the western edge of the central city an industrial zone (before 1966 called Léo-Ouest) flourishes near the site of the first depot established by the British-American explorer Sir Henry Morton Stanley. To its east lies the riverside residential and administrative district of Gombe, which houses most of the European population and the Congolese elite; the central government buildings and the embassy district are located there. The eastern sector (known before 1966 as Léo-Est), of which the wide Boulevard du 30-Juin forms the main artery, is a major commercial area. The waterfront, along Kinshasa's northern edge, is lined with quays and large warehouses. Ndolo, east of Gombe, comprises a complex of port facilities and industrial plants. The poorer areas extend southward on the east and west of Kinshasa. Among Kinshasa's satellite cities, Ndjili, to the southeast, has become a residential area, while Kimpoko upstream has been developed as an outer port. During the 1970s, wealthy businessmen and politicians built mansions, often of spectacular opulence, in Binza, an area in the western hills overlooking the city.

Government and residential section

There are a variety of architectural styles in Kinshasa. High-rise apartment blocks, luxuriously appointed banks, stores, and the offices of large corporations and government agencies characterize the centre of town. Some date from shortly after World War II, but the most prominent were constructed during the economic boom of the early 1970s. They include the Parliament, the president's palace (built by the Chinese), the headquarters of the national broadcasting corporation (Radio-Télévision Nationale Congolaise), the International Trade Center, the headquarters of the mineral marketing agency, and the unfinished tower dedicated to Nationalist leader Patrice Lumumba. Spacious villas surrounded by ornamental shrubs and flower gardens, and often also by high walls and iron bars, stand on tree-lined, paved boulevards that mark the elite residential districts. Dwellings in the less affluent communities often consist of tin-roofed, concrete-block houses and multiunit dwellings on unpaved streets, often inaccessible to vehicles. These in turn give way to hastily assembled shelters and rough pathways in the extensive squatting zones of the city, where many of the most recent immigrants reside.

Architectural landmarks

THE PEOPLE

The population of Kinshasa grew slowly at first (from 5,000 people in 1889 to 23,000 in 1923) but increased rapidly after 1940; after 1950 it doubled about every five years and had surpassed 4.5 million by the mid-1990s, a great many of whom lived in the squatting zones. Much of the population growth has been the result of Congolese migration and government expansion, but widening of the city's boundaries has caused some of the increase. Kinshasa has a young population. More than half the people are under 22 years of age, and only about 3 percent of the population is over 50.

Migration of peoples from the rural areas intensified greatly after independence as colonial restrictions were relaxed. Political troubles and the economic decline of rural areas and their lack of amenities and opportunities, as well as the attractions of the city, have contributed to this rural exodus, especially since the onset of civil strife in the 1990s. In its early years the city received immigrants from West Africa and the various neighbouring countries of Central Africa; since independence, however, most new inhabitants have come from within Congo, especially the nearby regions of Bandundu, to the west, and Bas-Congo (Lower Congo), to the south and east.

Population growth through migration

THE ECONOMY

Industry and commerce. Kinshasa is the most important consumer centre of the republic and the core of its industrial and commercial activity. The city serves as the headquarters of major public corporations and of privately owned industrial and commercial companies. It dominates the financial and commercial life of the republic and houses the head offices of the principal banks. Among Kinshasa's main industries are food processing and those producing consumer goods (*e.g.*, beer, textiles, and footwear), generally for domestic markets. Construction and various service activities also contribute to the city's economy. However, the political turmoil that has gripped the country since the downfall of the Zairean regime in 1997 has been debilitating for Kinshasa's economic activities.

The rapid expansion of Kinshasa's population has created serious problems in supplying the city with food, with the constant threat of shortages. The situation has been exacerbated by the country's economic woes since the late 1990s. In normal times, the busy central market is complemented by suburban markets lined with wooden stalls and by hawkers and street vendors selling in minute quantities to passers-by. The region of Bas-Congo supplies at least half of the food consumed in the capital. Other foodstuffs come from more distant regions of Congo or are imported. For those who can afford it, South Africa has been an important source of meat and fruit and vegetables, which are flown in. For the poor, however, Kinshasa is in some ways like an overgrown village, whose people forage at a considerable distance for firewood and keep gardens where they can find good soil. The demands of this vast urban population have caused extensive erosion in the surrounding countryside, as the soil is exhausted from overcultivation and trees cut for charcoal have not been replanted.

Transportation. Kinshasa's transportation system is inadequate in many respects. Economic problems and shortage of foreign exchange have caused severe deterioration, and there has been a continual crisis for lack of spare parts and replacement vehicles. Kinshasa is well served by roads, but its dense and rapidly increasing population causes much congestion. The city is connected by a paved road to Matadi, Congo's principal port, at the head of navigation on the Congo estuary, and by another to Kikwit, to the east. The railway line from Matadi, bypassing the rapids on the river below Kinshasa, brings in most of the country's imports, some of which are then conveyed upriver. The Congo is navigable to Kisangani, some 1,000 miles upstream, and a vast network of navigable stretches on its tributaries, connected by railways, brings almost all inland traffic carrying exports destined for Matadi down the Congo and through the port of Kinshasa. Ndjili International Airport, to the southeast, is one of Africa's largest air stations. A busy ferry connects the city to Brazzaville, the capital of the Republic of the Congo, across Malebo Pool. Within Kinshasa public transportation consists of grossly overcrowded buses, minibuses, taxis, and *fula-fula* (trucks adapted to carry passengers).

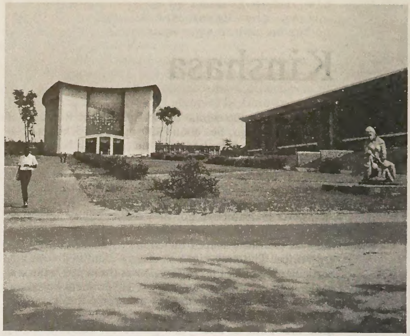
The port of Matadi

ADMINISTRATION AND SOCIAL CONDITIONS

Government and services. Kinshasa remains the place where all administrative decisions of importance are made; consequently, it is the centre of the nation's political life. The city houses the national government: the office of the president and the executive and legislative councils. Since 1982 the urban administration has consisted of a governor and two vice-governors, appointed by the president. They head the city council, consisting of the 24 zone commissioners appointed, also by the president, from among the councillors elected in each zone.

The administration is unable to adequately provide such services as running water, electricity, and sanitation throughout the city; town-planning and building-control agencies have had difficulty coping with the rapid growth of the city, much of which consequently lacks basic urban facilities. Some areas suffer from eroded housing lots and roadways, clogged open drains, and accumulated trash. The rate of violent crime, relatively low during the period of Zairean rule, has increased with the breakdown of central authority under the successor regime.

Health and education. Medical facilities, like other city services, are overwhelmed by the expansion of the population. The hospitals, clinics, and dispensaries of the public health system are insufficient in number and unevenly distributed, which, coupled with the problems of transportation, limits the health care they can provide for the public. The primary and secondary education system is similarly overextended, lacking sufficient facilities and teachers to cope with population growth. Institutions of higher education include the Université de Kinshasa (formerly Lovanium University), the largest of the country's three universities; a teacher-training college; a national school of administration and law; a school of telecommunications; and an academy of fine arts, as well as institutes of social research, political party indoctrination, medical training, and commerce. Kinshasa's School of Catholic Theology is internationally distinguished.



Campus of the Université de Kinshasa.
Michel Huet—HOA-QUI

Health and education problems

CULTURAL LIFE

Kinshasa is the dynamic centre of the nation's popular culture, of which the language is Lingala, the urban lingua franca. Congo's popular music is renowned throughout Africa; well-established bands from the country tour abroad to Europe and the Americas; popular music celebrities receive wide attention, and it is not unusual for their latest hit songs to give their names to fashions in women's dress materials, a medium of intense social competition. Like the popular songs, paintings sold on sidewalks express the social themes of the day. Daily papers and several periodicals are available to the populace. Television is an important medium of official communication to the people; it broadcasts news, speeches, a form of propaganda entertainment called "animation," popular bands, and occasional old European films. Radio and television broadcasting is in French, the official language, and local languages.

Radio and television

Modern Kinshasa has produced a considerable flowering of literature in novels, plays, and poetry by local writers. Painting and sculpture produced by artists of the Académie des Beaux-Arts are exhibited and sold at the academy. The collection of the Institut des Musées Nationaux is of great archaeological, ethnographic, and musicological as well as aesthetic interest, and it is of immense importance for scholars of traditional African art. Although traditional art of value may no longer be available in the city, workshops in the suburbs turn out imitations of masks and sculptures that represent all parts of Africa, as well as carved work in ivory and malachite. The city is known for some excellent restaurants and is the site of numerous nightclubs and motion-picture theatres.

History

The land on which Kinshasa grew was inhabited in ancient times, as were all the shores of Malebo Pool. The present city evolved from two villages, Nshasa and Ntamo (later known as Kintamo), dominated by the Bahumbu and frequented by Bateke fishermen and traders. Sir Henry Morton Stanley, on his visit in 1877, formed an alliance with the ruler of Kintamo, a wealthy ivory trader, and, despite French efforts to forestall him, was able to acquire a trading post site on his return in 1881. He named this post Léopoldville, for his patron, Léopold II, king of the Belgians. Although Stanley succeeded in opening river traffic as far north as Stanleyville (Kisangani since 1966) by portaging prefabricated steamers around the cataracts of the lower river, Léopoldville remained unimportant until the railway line from downstream Matadi was completed in 1898. A pipeline from Matadi to carry crude oil to the upriver steamers at Léopoldville was completed in 1914, and an air service was inaugurated between Léopoldville and Stanleyville in 1920. As a result, the administrative headquarters of the then Belgian Congo was transferred there from Boma in 1923.

Rail and pipelines spurring growth

As industries were established, residential zones grew up around them. In the 1930s the zones of Kinshasa, Barumbu, and Lingwala grew up near the port. After 1950 Lemba, Matete, and parts of Ndjili, to the southeast, were built to house the workers of the new industrial district of Limete, but the more centrally located communes (now zones) of Dendale (now Kasa-Vubu), Bandalungwa, and Ngiri-Ngiri became the social and political heart of the city. In 1960 Léopoldville became the capital of the new republic. Its name was changed to Kinshasa in 1966. The city prospered in the independence period, its population growing rapidly. In the mid-1990s Kinshasa became the focus of the rebel uprising against the Zairean regime of Mobutu Sese Seko, who was forced out of power in 1997. The successor regime, however, was also besieged by insurgents, and the city—overburdened with newcomers—entered a period of severe economic hardship.

BIBLIOGRAPHY. A full description of Kinshasa is contained in MARC PAIN, *Kinshasa: écologie et organisation urbaines*, 3 vol. (1978–79); the Republic of Zaire's *Atlas de Kinshasa*, issued by the INSTITUT GÉOGRAPHIQUE DU ZAÏRE, BUREAU D'ÉTUDES D'AMÉNAGEMENTS URBAINES (1975), contains an excellent descriptive text supporting the 44 thematic maps. Works of historical significance include HENRY M. STANLEY, *The Congo and the Founding of Its Free State*, 2 vol. (1885, reprinted 1970); and CRAWFORD YOUNG, *Politics in the Congo: Decolonization and Independence* (1965). JEAN S. LA FONTAINE, *City Politics: A Study of Leopoldville, 1962–63* (1970), describes the city's social organization; and MBUMBA NGIMBI, *Kinshasa, 1881–1981: 100 ans après Stanley: problèmes et avenir d'une ville* (1982), describes its administrative history and problems. KANKUENDA M'BAYA, *Les Industries du pôle de Kinshasa* (1977), discusses the city's economic role.

(J.O.A./W.MacG./J.MacG./Ed.)

Korea

Extending southward from the Chinese historic area of Manchuria (now the Northeast, Chinese provinces of Liaoning, Kirin, and Heilungkiang) and Siberia (Asian Russia) on the northeastern Asian mainland, the peninsula of Korea is some 600 miles (970 kilometres) long and from 125 to 200 miles wide. The elongated and irregularly shaped peninsula serves to divide the Yellow Sea to the west from the East Sea (Korean: Tonghae), or Sea of Japan, to the east. In the southeast the peninsula is separated from the Japanese island of Tsushima by the Korea Strait.

Korea is largely a mountainous land. The highest elevations on the peninsula are found in the northern interior, while the drainage divide of central Korea lies in the rugged, mountainous terrain along the eastern side. The relief of the western and southern parts of the peninsula is considerably lower than that of the north and east. While the eastern coastline generally is sharply defined, the southern and western coastlines are deeply indented; most of the peninsula's approximately 3,500 islands are found in the south and west. Korea has no active volcanoes, and earthquakes are rare.

The Korean people share a common ethnic origin with other peoples of North Asia, and the Korean language belongs to the Altaic language family of the region. Korea is strategically located between China, Japan, and the Siberian Far East region, and it long has suffered from the inroads of aggressive neighbours. China was antagonistic toward the peoples on its northern borders, with whom the Koreans were associated, and assumed an aggressive attitude toward the Koreans.

After the unification of the Korean peninsula by the kingdom of Silla (7th century AD), Korea maintained friendly relations with China out of the necessity to deter invasion by the northern nomadic peoples. Later Manchuria and Mongolia came under the successive control of the northern nomadic peoples and served as bases for invasion of Korea and China. To cope with this threat, Korea and China were pressed to form a military alliance.

Cultural homogeneity between Korea and China also promoted their friendly relations. China's culture was regarded as more advanced than Korea's, and Korea absorbed much from China—*e.g.*, its writing system, law, Confucianism, and fine art. Nevertheless, Korea had its own language and invented its own alphabet; it formed a distinct way of life and adapted and improved upon all borrowed culture to fit indigenous Korean needs.

By the late 19th century, Western influence began to be felt in East Asia. Japan took the lead in opening its doors to the West. Korea also attempted to institute reforms but in the process fell prey to Japanese imperialism. Japan occupied Korea in 1905 and ruled it as a colony from 1910 to the end of World War II (1945).

On liberation from Japanese rule, Korea was divided by the Allies for the purpose of accepting the Japanese surrender. This division, a line at latitude 38° N (commonly called the 38th parallel), was perpetuated by the U.S. occupation of the south and the Soviet occupation of the north. In 1948 the Republic of Korea (South Korea) was formed in the south, with its capital at Seoul (Sŏul), and the Democratic People's Republic of Korea (North Korea) was proclaimed in the north, with its capital at P'yŏngyang. Following the Korean War (1950–53), an armistice was signed in 1953 that has left the peninsula divided by a demilitarized zone that cuts across the land from north of the 38th parallel in the east to south of it in the west.

The name by which Korea is best known to its own people is Chosŏn, which may be translated as "Land of the Morning Freshness." South Korea uses Taehan ("Great Han," Han being another name for Korea) as its official name, while North Korea continues to use the name Chosŏn. The Western name, Korea, was derived from the Koryŏ dynasty (AD 918 to 1392) and may be literally translated as "High and Beautiful."

This article first discusses the history of Korea from ancient times to the political division of the peninsula after World War II, followed by a discussion of each of the modern Korean states.

The article is divided into the following sections:

HISTORY

Korea to *c.* 1400

THE DAWN OF HISTORY

Archaeological, linguistic, and legendary sources support the view that the Korean peninsula was settled by Tungusic-speaking peoples who migrated in waves from Manchuria and Siberia. They settled along the coasts and moved up the river valleys. These peoples formed the dominant ethnic stock of the Korean people and developed the Korean language. There was a close relationship between Korean culture and that of neighbouring peoples in the Neolithic Period and the Bronze Age. For example, Korean combware pottery, widely used in the Neolithic Period, is commonly found in northeastern Asia; Korean bronze daggers, belt hooks, and knobbed mirrors also display the traits of bronze tools unearthed in the region.

The Stone Age. Stone artifacts of the Paleolithic Period have been unearthed at Kulp'o-ri in North Hamgyŏng province (North Korea) and at Sŏkchang-ni in South Ch'ungch'ŏng province (South Korea). Of 13 stratified Paleolithic sites, each cultural stratum produced chipped-stone tools of different shapes. Carbon-14 dating indicates the Paleolithic provenance of Sŏkchang-ni. Dwelling sites with round fireplaces were discovered there along with carved pebbles.

Neolithic life

The Neolithic Period was well established by 3000 BC. A major characteristic was the use of combware pottery, chiefly found at seashore and river-basin sites, where inhabited places and shell mounds also have been discovered. In addition, stone spears and flint arrowheads have been found, as well as bone hooks and stone weights used for fishing. Remains of the Late Neolithic Period include stone plows and sickles, which indicate the beginning of farming. People lived in dugouts, mostly shallow round or rectangular hollows with fireplaces in the centre that may have been covered with thatched roofs. These shelters were huddled together in groups. The size of such villages is yet to be determined, but legends indicate the family members lived together, forming clan communities.

The use of metals and the emergence of tribal states. Bronze ware was probably first used about the 8th century BC, though some scholars surmise that it predates the 10th century. As the Bronze Age started, the design of pottery changed to undecorated earthenware. The uncovering of such pottery indicates that Bronze Age Korean people lived on hillsides, in dugouts raised slightly aboveground. Half-moon-shaped stone reaping knives and grooved stone axes used for hoeing show that rice farming was practiced, and bronze daggers and bronze arrowheads indicate participation in wars of conquest. Dolmens, used as tombs, which were discovered in southern Manchuria and the Korean peninsula, show the boundary of ancient Korean culture. Since only prominent persons were buried in dolmens, their number and location indicate that many small Bronze Age tribal states were probably established by powerful men.

The most advanced state was Old Chosŏn, established in the Taedong River basin. According to legend, the son of heaven, Hwanung, descended to earth and married a bear-turned-woman, who bore a son, Tan'gun, the founder of Chosŏn. Perhaps Tan'gun and his descendants ruled a tribal state in which rituals and politics were not separated.

Chosŏn developed into a league of tribes in the area of the Taedong and Liao rivers (*c.* 4th century BC). About this time ironware came to be used. Iron plows and sickles indicate the use of animals in farming and more efficient harvesting methods. Wooden houses were built on the ground, and *ondol,* a floor-heating device, was developed. The appearance of iron weapons, horse equipment, and coaches indicates that horses and chariots were employed in wars. Wiman (Wei Man in Chinese), said to have defected from China, became ruler of Chosŏn about 194 BC. More likely, he was indigenous to Chosŏn. Wiman's Chosŏn was overthrown by the Han empire of China and replaced by four Chinese colonies in 108 BC.

THE DEVELOPMENT OF ANCIENT STATES

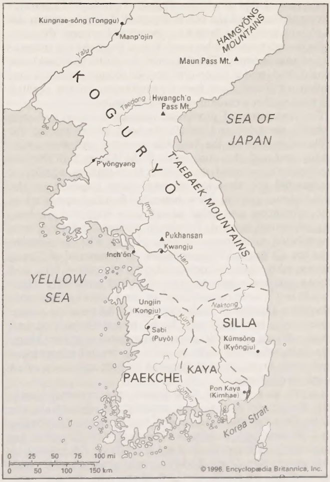
The Three Kingdoms. Apart from Chosŏn, the region of Korea developed into tribal states. To the north, Puyŏ rose in the Sungari River basin of Manchuria. Chin, which had emerged south of the Han River in the 2nd century BC, was split into three tribal states—Mahan, Chinhan, and Pyŏnhan. These states formed leagues, or tribal federations, centred on a leading state. The tribal leagues stretched across a wide area from the Sungari basin in Manchuria to the southern Korean peninsula. They evolved into three rival kingdoms—Koguryŏ, Paekche, and Silla. According to legends, Koguryŏ was founded by Chu-mong in 37 BC, Paekche by Onjo in 18 BC, and Silla by Pak Hyŏkkŏse in 57 BC. The actual task of state building, however, was begun for Koguryŏ by King T'aejo (reigned AD 53–146?), for Paekche by King Koi (reigned 234–286), and for Silla by King Naemul (reigned 356–402).

Common characteristics of the Three Kingdoms

The Three Kingdoms shared several common characteristics. They evolved into statehood through frequent wars of expansion, centralized military systems were organized, and training institutions (*kyŏngdang* in Koguryŏ, *hwarangdo* in Silla) were developed. The power of the king in each state was strengthened, and hereditary monarchies evolved.

Another common characteristic was the appearance of powerful aristocracies composed of tribal chiefs who moved to the capital. The aristocrats were divided into several social classes with certain privileges as they advanced socially and politically. Silla's *kolp'um* ("bone-rank") system, in which the families of rulers customarily monopolized political power, was typical. Silla had a state deliberative body, the Council of Nobles (Hwabaek), which made important decisions. The council's membership consisted of men of *chin'gol* ("true-bone") class, who were of the high aristocracy.

The kingdoms all achieved a centralization of power. Each one was divided into administrative units—the largest called *pu* in Koguryŏ, *pang* in Paekche, *chu* in Silla—that controlled many castles. To these provincial



Korea during the Three Kingdoms period (*c.* AD 400).

units the central government sent officials who made sure that the people, as royal subjects, provided taxes and corvée labour.

The Three Kingdoms developed highly sophisticated cultures. Each compiled its own history, apparently to consolidate the authority of the state. Also noteworthy was the introduction of Buddhism, which was regarded at the time as the state religion for the protection and welfare of the state.

Unified Silla. With the support of China, Silla conquered and subjugated Paekche in 660 and Koguryŏ in 668. Not until 676 did Silla drive out the Chinese and gain complete control of the Korean peninsula. The surviving Koguryŏ people in northern Manchuria established Parhae (or Palhe; P'o-hai in Chinese), under the leadership of Tae Cho-yŏng, which soon came into direct confrontation with Silla. This period may be called an age of separate southern and northern states; it is customary, however, for historians to place the primary focus on Silla because little is known about Parhae, though it grew into a highly civilized state that the Chinese called the "Prosperous Country of the East." After Parhae's demise its territory fell under the control of the northern nomadic peoples and has not since been a part of Korean history.

Unified Silla saw the maturing of an absolute monarchy, which effectively eliminated the influence of the Council of Nobles. A central administrative body called the Chancellery (Chipsabu) was established to enforce royal decrees. Aristocrats were now granted salaries and land, but the latter was to revert to the state when the aristocrats left office. Thus, the aristocracy's direct control over land and the populace was reduced. Monarchs built extravagant palaces and royal tombs at Kŭmsŏng (modern Kyŏngju, S.Kor.), the Silla capital. The state was divided into administrative units by province (*chu*), prefecture (*kun*), and county (*hyŏn*). Five provincial capitals prospered as cultural centres.

Buddhism

Avataṃsaka Buddhism provided the ideological backing for autocratic monarchy and the aristocracy. The underprivileged populace was attracted to the *Sukhāvatī-vyūha-sūtra* of Pure Land Buddhism, which promised bliss in the next world. The legacy of Silla Buddhism can be seen in many beautiful temples and great works of art, the most remarkable of which—Pulguk Temple, Sŏkkuram (a grotto shrine), and the bell at Pongdŏk Temple—are in the Kyŏngju area. Confucianism prospered among the low-echelon aristocrats, who used it as a foothold for bureaucratic advancement. The National Academy (Kukhak) was established, and a proto-civil service examination system, called *toksŏ samp'um kwa* ("examination in the reading of texts in three gradations"), was installed.

The emergence of provincial magnates. Frequent succession struggles and rebellions took place among the Silla aristocrats in the late 8th century, and they eventually restored the authority of the Council of Nobles and overthrew royal despotism. Low-ranking aristocrats demanded the abolition of restrictions imposed by the strict status system. New powerful families appeared in the provinces, and their power grew with the weakening of central control. Provincial military fortresses were established to suppress Chinese pirates. The most active was the Ch'ŏnghae fortress under the command of Chang Po-go, who virtually monopolized trade with China and Japan and had a private navy of 10,000 men. Silla settlements in Chinese coastal cities in the Shantung Peninsula also were engaged in trade. Also powerful were the village rulers, who became "castle lords" by establishing control over military, administrative, and economic affairs. Many peasants, who were taxed by both the central government and castle lords, chose to become drifters or thieves, often staging rebellions.

Largely as a result of these trends, two provincial leaders, Kyŏnhwŏn and Kungye, established, respectively, the Later Paekche (892) and Later Koguryŏ (also called Majin or T'aebong; 901) kingdoms. Together with Silla, they are commonly referred to as the Later Three Kingdoms. In this period Sŏn (Zen) Buddhism was most popular, with its emphasis on the importance of realizing, through contemplation, the inborn Buddha nature of the individual.

KORYŎ

Social structure and culture. Koryŏ was founded in 918 at Songak (modern Kaesŏng, N.Kor.) by Wang Kŏn, who in 936 established a unified kingdom in the Korean peninsula. Wang Kŏn went to great lengths to absorb the people of the overthrown states, even accepting the survivors of Parhae, which had been destroyed by the Khitan (Liao). Proclaiming itself the successor of Koguryŏ, Koryŏ launched active campaigns to recover lost territory and clashed frequently with the Khitan in the north. Koryŏ eventually expanded its territory to the Yalu (Korean: Amnok) River.

The Koryŏ ruling class consisted largely of provincial castle lords and former Silla aristocrats. The rulers held their family lineage in high esteem. Marriage into a powerful family, especially a family of royal blood, was an important means for maintaining and elevating one's social and political status. Sons of a family above the fifth of nine official grades received official posts without undergoing civil service examinations.

Government structure

The central government consisted of two supreme organs: the Three Chancelleries (Samsŏng) and the Royal Secretariat (Chungch'uwŏn). These two formed the Supreme Council of State. Koryŏ politics was thus centred in the aristocratic council. Officials above the fifth grade were given land for permanent possession. Even the land supposed to be returned was actually handed down for generations because the grantees' sons usually became officials. Land was the primary source of wealth, and aristocrats expanded their landholdings by reclamation, purchase, or seizure.

The aristocracy embraced Buddhism as the religion for spiritual fulfillment and personal happiness and Confucianism for its political precepts and ethical principles. The same was true of the government, which built grand Buddhist temples, such as Hŭngwang Temple, to observe rituals and pray for the prosperity of the nation but which also set up a national academy, Kukchagam, to inculcate Confucianism.

Military rule. Civil officials constituted the core of the ruling class, with the military generally subjected to discrimination. Even the supreme commander for military affairs was a civilian. Military officials were not eligible for the second grade or above in the official hierarchy and were excluded from the Supreme Council of State. Even in the same official grade, military men received less land than did their civilian counterparts. This discrimination eventually led to a military coup d'état in 1170. The rebels massacred a large number of civil officials and seized complete control of government. The struggle for hegemony that erupted among the leaders was won by General Ch'oe Ch'ung-hŏn, who established a military regime of his own that lasted roughly 60 years.

The monarch remained as a figurehead, deprived of political power, which was in the hands of the Ch'oe family. The Ch'oe had a private army for personal protection and a new public military organization for national security. The latter also served, in effect, as their private army. The Ch'oe also established a body of civilian officials to manage the state's personnel administration, thus controlling both the military and civil branches of government.

Buddhism was suppressed, and many monks retreated to remote mountain areas. There they formed a new Sŏn sect called Chogye, which became the mainstream of Korean Buddhism. The underprivileged peasantry, stimulated by a general political atmosphere in which subordinates rose against superiors, staged rebellions across the country over a period of 30 years. The upheavals were at first a natural and spontaneous protest against oppression, but they developed into an organized struggle for emancipation and for power. The struggle, eventually brought under control through appeasement and by the use of force, was nevertheless instrumental in improving the lot of the peasantry.

In 1231 the Mongols invaded Koryŏ, and the Ch'oe regime resisted them for nearly 30 years. Even peasants and servants stood up bravely. The Mongols, who had conquered most of Eurasia, found it difficult to take Koryŏ by force. As the exploitation of the peasantry by the Ch'oe grew more severe, however, the people became estranged, and the regime was finally overthrown by civilian leaders, who in 1258 concluded a peace treaty with the invaders.

Treaty with the Mongols

Social change in later Koryŏ. After the peace treaty, Koryŏ was subject to occasional political interference from the Mongols but retained its political and cultural identity. Koryŏ went to some lengths to show its national and cultural superiority over the invaders by producing highly refined poetry and works on national history.

During this period, large manors operated by powerful aristocrats were created throughout the country. The landowners lived in the capital and sent private retainers and servants to collect taxes from the commoners who tilled their land. The tenants often were forced to pay taxes to more than one owner because landholders shared ownership. Tenants were also subject to forced labour and military duty for the state. Many peasants chose to become serfs (*nobi*) in order to seek protection by aristocrats and to avoid the state levies. Some aristocrats captured drifters and illegally made them serfs. These serfs were not slaves in the Western sense but were actually on a level with tenants. The increase in the number of landholders and serfs resulted in a reduction of state tax revenue and of the number of people available to be mobilized in war.

Through the civil service examination, the central government recruited a new bureaucratic force consisting of scholar-officials (*sadaebu*), who generally had small farms under their own management in their native districts. These men held Buddhism in disdain and were not satisfied with superficial interpretations of the Chinese Classics. They adopted Neo-Confucianism, which introduced a metaphysical approach to the understanding of the universe. Inadequate government resources precluded the granting of land to newly appointed officials commensurate to their rank, resulting in a demand for land reform. Eventually, with the support of General Yi Sŏng-gye, the disgruntled scholar-officials seized power and established a new land-distribution system, under which land was granted according to the rank of office. These reforms spelled the end of the Koryŏ dynasty in 1392 and ushered in the new Chosŏn (Yi) dynasty. (K.-b.L./Y.I.L.)

Korea since *c.* 1400

THE CHOSŎN (YI) DYNASTY

The establishment of a Confucian state. When the dynasty was established, the territory under its control was named Chosŏn, with the approval of the emperor of China. The Chosŏn dynasty, with 26 monarchs, ruled from 1392 until the Japanese annexation of Korea in 1910. Hanyang (now Seoul) was made the capital. The Confucian ethical system was officially adopted and replaced Buddhism, which had become corrupt. Chosŏn society was dominated by a hereditary aristocratic class, the *yangban* (literally, "two orders," meaning civil and military officials). Members of the *yangban* devoted themselves to the study of Neo-Confucian orthodoxy and, through civil service examinations, held public offices, their sole profession. Since they controlled all aspects of Chosŏn society and owned most of the land, the Chosŏn dynasty can be described as a *yangban* society.

Sejong the Great

Early *yangban* society flourished intellectually and culturally, especially during the reign of Sejong the Great, the fourth monarch. With the technique of movable-type printing, developed in Korea in 1234, many publications were produced in such fields as medicine, astronomy, geography, history, and agriculture. In 1420 a royal academy called the Hall of Worthies (Chiphyŏnjŏn) was established, where bright young scholars engaged in study and research. In 1443 the Korean phonetic alphabet, Hangul (Korean: *han'gŭl*), was completed under Sejong's direction.

In the reign of Sejo, the seventh monarch, a powerful centralized and *yangban*-oriented government structure emerged. The country was divided into eight administrative provinces, and all officials were appointed by the central government. Laws were codified, and the highest administrative body was the State Council.

Late in the 15th century Korean scholars made original contributions to the theoretical refinement of Confucianism. In the mid-16th century many of these scholars were recruited into government service. Idealistic in orientation, they criticized the bureaucratic establishment and recommended drastic measures for the realization of Confucian ideals. But relentless counterattacks and pressures forced most of the scholars to retire from their posts, whereupon they established private academies called *sŏwŏn.* These academies produced many eminent scholars, including Yi Hwang (T'oegye) and Yi I (Yulgok), whose distinct theories of the universe evolved into rival schools of thought.

Foreign invasions. In 1592 Toyotomi Hideyoshi, the Japanese military leader who had just reunified Japan, sent a large force to Korea in an alleged attempt to invade China. The Korean land forces suffered a series of defeats, but Korean naval forces, led by Admiral Yi Sun-shin, secured full control of the sea. Yi won the greatest naval victories in Korean history, over the Japanese squadrons off Korea's southern coast. The national crisis brought people of almost all ranks, including Buddhist monks, to volunteer in fighting the Japanese. Ming China also dispatched troops to aid Korea. After one year the Japanese were forced to retreat, although another invasion followed in 1597. After Toyotomi's sudden death in 1598 the Japanese withdrew. The war left most of Korea in ruins. Palaces, public buildings, and private homes were burned, and many cultural treasures were lost or destroyed. Scholars and artisans were kidnapped to Japan, where they were forced to teach Korea's advanced technology.

In the early 17th century nomadic Manchu violated the borders of both Ming China and Korea. Ming and Korean punitive attacks on Manchu strongholds in 1619 were beaten back, and in 1627 the Manchu overran northern Korea. Only after Korea had agreed to recognize their demand for "brotherhood" did the Manchu withdraw from the occupied territory. In 1636 the Manchu captured Seoul and wrested an unconditional surrender from the king. The Manchu then overthrew the Ming and in 1644 established the Ch'ing dynasty; the tribute that Korea had paid to the Ming was switched to the Ch'ing.

***Silhak* and popular culture.** A series of significant changes in Korea began in the mid-17th century and made a great impact on virtually every sector of Korean society in the 18th century. In agriculture, rice transplantation became popular, and irrigation systems were improved. Advances in farming resulted in dramatic increases in agricultural production and raised the standard of living for peasants. With the cultivation of such special crops as tobacco and ginseng, commerce and trade developed apace. The government started minting coins and collecting farm rent in cash. Markets were held in many places across the country. Particularly active were merchants from Kaesŏng, who established a national network that put every fair in the country within their sphere of influence.

The rise of markets

In the realm of scholarship, attention shifted from speculative theorizing to matters of practical relevance—the needs of society and state. Scholars who engaged in such studies are identified with the *silhak* (*sirhak*), or "practical learning," school. They fell into four major groups. One group advocated comprehensive administrative reform, calling upon the government to rationalize the systems of civil service examination, education, taxation, and land administration. Another group stressed the need to foster commerce, industry, and technology. A third conducted critical examinations of the Confucian Classics, while the fourth focused on the study of Korean history, geography, and language.

Comparable new trends appeared in arts and letters. Popular literary and artistic works came into fashion—a marked change from the tradition of catering exclusively to the upper class. The new works not only were written in the easy-to-read Hangul but also gave frank expression to popular discontent. Singing dramas or traditional Korean operas, called *p'ansori,* most of them adapted from vernacular novels, were also popular with the masses. Many artists specialized in pictures of blacksmiths at work, peasants in the field, traditional wrestling matches, and rural landscapes. Pottery with a simple blue and white glaze was produced in large quantities for popular consumption.

The introduction of Roman Catholicism. Significant numbers of Europeans began to arrive in East Asia in the mid-16th century. In 1656 a Dutch merchant ship went aground off the southern shore of Cheju Island, and its 36 surviving crewmen were taken to Seoul for detention. Thirteen years later Hendrik Hamel and seven others escaped and returned home. Hamel wrote an account of his experiences—the first book on Korea published in Europe.

Along with the European merchants came Roman Catholic priests. Korea's first significant contact with Christianity was through missionaries in China. Korean envoys to China in the 16th century brought back with them a world atlas and scientific instruments made by the priests, as well as literature on science and Christianity. Some *silhak* scholars had converted to Catholicism by the late 18th century, even before missionaries reached Korea. Most of the early converts were scholars of aristocratic background. Commoners were later attracted to Catholicism, finding hope in the Christian doctrine of equality of all people before God and a new source of solace in the Christian belief in life after death. Catholicism spread from Seoul to the provinces steadily.

The incompatibility of Catholicism with Confucianism posed a serious problem. The two could not compromise on the great importance Confucianism attached to ancestor worship, which Catholicism rejected as flagrant idolatry. The government began to suppress the Catholic church in the belief that it defied the existing sacrosanct mores of Confucianism. During persecutions in 1801, 1839, and 1866, scholar-converts were either put to death or forced to apostatize; foreign missionaries were ferreted out and beheaded. But rank-and-file Catholics rallied around the church, and it was precariously maintained. In 1831 the Holy See set up a Korean parish, and French priests smuggled themselves into the country to engage in clandestine proselytism.

Persecutions

The advent of *silhak,* popular arts, and Roman Catholicism in the 17th and 18th centuries indicates a modern Korea in the making. But in the 19th century boy kings came to the throne in succession, and their maternal relatives seized power and plunged the government into a state of chaos. One peasant uprising followed another in the provinces, and the whole nation seethed with popular discontent and resentment.

Many peasants sought refuge in religion. A new religion founded in 1860 by Ch'oe Che-u, a fallen country *yangban* scholar, advocated sweeping social reform. It had much in common with traditional animism and appealed to the peasantry. This religion was called Tonghak, or "Eastern Learning," as a counterpoise to Sŏhak, or "Western Learning"—*i.e.,* Catholicism.

CONTACT WITH WORLD POWERS

Opening the door. King Kojong was too young to rule when he ascended the throne in 1864, and his father, Yi Ha-ŭng, known as the Taewŏn'gun ("Prince of the Great Court") became the de facto ruler. The Taewŏn'gun set out to restore the powers of the monarchy and pursued a policy of national exclusionism. He put into force bold political reforms, such as faction-free recruitment of officials and the closing of many private Confucian academies.

During his rule, Western men-of-war and merchant vessels came in search of trade and friendship, but the Taewŏn'gun refused them. Korean soldiers and civilians burned and sank the American merchant ship *General Sherman* at P'yŏngyang in retaliation for lawless acts committed by the crew. Koreans repulsed two attacks by French warships in 1866. In 1871 an American flotilla came to obtain a shipwreck convention but, encountering Korean resistance, left. Such incidents strengthened the Taewŏn'gun's resolve to keep the country's doors closed.

Japan repeatedly made futile attempts to establish diplomatic relations with Korea. The Japanese militarists thereupon raised an outcry for a war of conquest on Korea. Meanwhile, the Taewŏn'gun came under widespread criticism for the enormous financial burden he had imposed on the people. He relinquished his power in 1873 in favour of Kojong. Queen Min and her relatives took over the helm of state and initiated policies opposed to those of the Taewŏn'gun. Japan, which had been watching developments in Korea, dispatched a squadron of warships and pressured Korea to sign a treaty of commerce and friendship. The ports of Pusan, Wŏnsan, and Inch'ŏn were subsequently opened to Japanese trade.

Treaty with Japan

The growing Japanese presence in Korea was disturbing to the rulers of Ch'ing China. When conservative soldiers tried to restore the Taewŏn'gun, the Ch'ing used it as a pretext for stationing troops in Korea. Thus began a period of aggressive Chinese interference in Korean affairs. China forced Korea to sign a trade agreement that heavily favoured Chinese merchants. Korea signed a treaty of commerce and friendship with the United States (1882) through the good offices of China. Similar treaties with Great Britain, Germany, Russia, and France followed, and foreign missions were established in Seoul.

Once the doors were opened, a modernization movement began. Students and officials were sent to Japan and China; Western-style schools and newspapers were founded. The government, however, could not proceed with a consistent policy of modernization, for the king was feebleminded and the ruling class was divided into radicals and moderates.

In 1884 the radicals seized power in a coup d'état and formulated a bold blueprint for reform. Chinese troops, however, moved in and overthrew their three-day-old regime. This led in 1885 to the signing of the Li-Itō Convention, designed to guarantee a Sino-Japanese balance of power on the Korean peninsula.

The Tonghak Uprising and government reform. Government expenditures greatly increased, largely because of appropriations for machinery imports and government reorganization, and the difficult financial situation was aggravated by obligations to pay reparations to Japan. Heavier tax levies were imposed on peasants, who provided the bulk of government revenue. The import of such necessities as cotton textiles upset the traditional self-sufficiency of the farming community. Furthermore, usurious loans by Japanese rice dealers contributed to reducing the peasantry to abject poverty. Angry peasants turned increasingly to Tonghak.

Despite ruthless government persecution, Tonghak took deep root in the peasantry. Its followers staged large-scale demonstrations calling for an end to injustice. A negative official response precipitated the Tonghak Uprising (1894), in which the Tonghak followers and the peasantry formed a united front to demand reform. Government troops armed with Western weapons suffered ignominious defeats in the southern provinces, weakening the government's military grip on the country. Foreign intervention seemed the last resort open to the rulers, and Chinese troops soon moved in at the request of the government. Simultaneously, Japan, without invitation, dispatched a large military contingent, and the two foreign powers were in sharp and sudden confrontation.

The rebels laid down their arms voluntarily to defuse the threat, but the Sino-Japanese War broke out in July 1894. Japan emerged victorious, and the two belligerents signed the Treaty of Shimonoseki in April 1895, which recognized Japanese hegemony in Korea.

At the instigation of the Japanese, the Korean government initiated a wide range of reforms during the war. It set up a Deliberative Council to undertake reforms, issued pertinent decrees, and formed Western-style institutions and a cabinet. Civil service examinations were discontinued, and such social practices as class discrimination were abolished. Public reaction to the reforms was unfavourable. The government realized that old customs and institutions die hard and that reform takes more than mere decrees and imitation of things Western.

The international power struggle and Korea's resistance. Japan's supremacy in Korea and its subsequent acquisition of the Liaotung Peninsula in Manchuria were more than Russia, with its long-cherished dream of southward expansion in East Asia, could tolerate. With German and French support, Russia pressured Japan to return the peninsula to China. At the same time, encouraged by Russia, the Korean government began to take an anti-Japanese course. The Japanese thereupon engineered the assassination of Queen Min (October 1895), the suspected mastermind behind the anti-Japanese stance. Fearing for

Russian intervention

his own life, King Kojong took refuge in the Russian legation, where he granted such concessions as mining and lumbering franchises to Russia and other powers.

A popular movement for the restoration of Korean sovereignty arose under the leadership of such figures as Sŏ Chae-p'il (Philip Jaisohn). Returning from many years of exile, Sŏ organized in 1896 a political organization called the Independence Club (Tongnip Hyŏphoe). On the urging of the Tongnip Hyŏphoe, the king returned to his palace and declared himself emperor and his kingdom the Great Korean (Tae Han) Empire.

The Boxer Rebellion in China (1900) led to a Russian invasion of Manchuria and to the Russo-Japanese War (1904–05). The Korean government at first declared neutrality, but under Japanese pressure it signed an agreement allowing Japan to use much of its territory for military operations against the Russians.

Japan was the victor, and the resulting Treaty of Portsmouth (September 1905), signed through the mediation of the United States, recognized Japan's undisputed supremacy in Korea. Its hand thus strengthened, Japan forced the Korean emperor into signing a treaty that made Korea a Japanese protectorate (November 1905).

Although the Korean emperor sent a secret emissary to the international peace conference held at The Hague in 1906 to urge the great powers to intercede with Japan on behalf of Korea, the mission failed, serving only to infuriate Japan. Under Japanese coercion, Emperor Kojong then abdicated in favour of his son, Emperor Sunjong. The Korean army was disbanded, and in 1910 Japan annexed Korea.

Japanese annexation of Korea

A section of the Korean army led by deposed officials and Confucian scholars took up arms against the Japanese in the southern provinces following the 1905 treaty. For five years anti-Japanese guerrilla units, called the "righteous armies," effectively harassed the Japanese occupation forces, especially in 1908–09. With the annexation, however, they were driven into Manchuria. Large numbers of Koreans emigrated to Manchuria, Siberia, and Hawaii before and after 1910.

KOREA UNDER JAPANESE RULE

Military control. Japan set up a government in Korea with the governor-generalship filled by generals or admirals appointed by the Japanese emperor. The Koreans were deprived of freedom of assembly, association, the press, and speech. Many private schools were closed because they did not meet certain arbitrary standards. The colonial authorities used their own school system as a tool for assimilating Korea to Japan, placing primary emphasis on teaching the Japanese language and excluding from the educational curriculum such subjects as Korean language and history. The Japanese built nationwide transportation and communications networks and established a new monetary and financial system. They also promoted Japanese commerce in Korea while barring Koreans from similar activities.

The colonial government promulgated a land-survey ordinance that forced landowners to report the size and area of their land. By failing to do this, many farmers were deprived of their land. Land and forestry owned jointly by a village or a clan were likewise expropriated by the Japanese, since no single individual could claim them. Much of the land thus expropriated was then sold cheaply to Japanese. Many of the dispossessed took to the woods and subsisted by slash-and-burn tillage, while others emigrated to Manchuria and Japan in search of jobs.

The March 1st Movement. A turning point in Korea's resistance movement came on March 1, 1919, when nationwide anti-Japanese rallies were staged. The former emperor, Kojong, the supreme symbol of independence, had died a few weeks earlier, bringing mourners from all parts of the country to the capital for his funeral. A Korean Declaration of Independence was read at a rally in Seoul on March 1. Some 2,000,000 students and citizens took to the streets, demanding independence. This movement took the form of peaceful demonstrations, appealing to the conscience of the Japanese. The Japanese, however, responded with brutal repression. They arrested some 47,000 Koreans, of whom about 10,500 were indicted, while some 7,500 were killed and 16,000 wounded.

In September independence leaders, including Yi Tong-nyŏng and An Ch'ang-ho, who in April had formed a Korean provisional government in Shanghai, elected Syngman Rhee as president. It brought together all Korean exiles and established an efficient liaison with leaders inside Korea. Japan realized that its iron rule required more sophisticated methods. The gendarmerie gave way to an ordinary constabulary force, and partial freedom of the press was granted. But the oppressive and exploitative Japanese colonial policy remained ruthless, though using less conspicuous methods.

Taking advantage of a wartime business boom, Japan took leaps forward as a capitalist country. Korea became not only a market for Japanese goods but also a fertile region for capital investment. Meanwhile, industrial development in Japan was achieved at the sacrifice of agricultural production, creating a chronic shortage of rice. The colonial government undertook projects for increasing rice production throughout Korea. Many peasants were ordered to turn their dry fields into paddies. The program was temporarily suspended during the worldwide economic depression in the early 1930s. It soon resumed, however, in order to meet the increased needs of the Japanese military in its war against China, which began in 1931. Most Koreans were forced to subsist on low-quality cereals imported from Manchuria instead of their own rice.

Increase in rice production

The end of Japanese rule. A major anti-Japanese mass rally was held in Seoul in 1926, on the occasion of the funeral of Emperor Sunjong. A nationwide student uprising originated in Kwangju in November 1929, demanding an end to Japanese discrimination. These and other resistance movements were led by a wide spectrum of Korean intellectuals.

In 1931 the Japanese imposed military rule once again. After the outbreak of the Sino-Japanese War (1937) and of World War II in the Pacific (1941), Japan attempted to obliterate Korea as a nation: Koreans were forced to worship at Japanese Shintō shrines and even to adopt Japanese-style names, and academic societies devoted to Korean studies as well as newspapers and magazines published in Korean were banned. The Japanese desperately needed additional manpower to replenish the dwindling ranks of their military and labour forces. As a consequence, hundreds of thousands of able-bodied Koreans were drafted to fight for Japan and to work in mines, factories, and military bases.

When Shanghai fell to the Japanese, the Korean provisional government moved to Chungking in southwestern China. It declared war against Japan in December 1941 and organized the Korean Restoration Army, composed of independence fighters in China. This army fought with the Allied forces in China until the Japanese surrender in August 1945, which ended 36 years of Japanese rule over Korea. (K.-r.L./Y.I.L.)

DIVISION OF KOREA

The Cairo Declaration, issued on Dec. 1, 1943, by the United States, Great Britain, and China, pledged independence for Korea "in due course." This vague phrase aroused the leaders of the Korean provisional government in Chungking to request interpretation from the United States. Their request, however, received no answer. At the Yalta Conference held in February 1945, U.S. President Franklin D. Roosevelt proposed to Soviet Premier Joseph Stalin a four-power trusteeship for Korea consisting of the United States, Great Britain, the U.S.S.R., and the Republic of China. Stalin agreed to Roosevelt's suggestion in principle, but they did not reach any formal agreement on the future status of Korea, and after the Yalta meeting there was a growing uneasiness between the Anglo-American allies and the U.S.S.R.

Throughout the Potsdam Conference in July 1945, U.S. military leaders insisted on encouraging Soviet entry into the war against Japan. The ensuing Potsdam Declaration included the statement that "the terms of the Cairo Declaration," which promised Korea its independence, "shall

Initial concord on Korea

be carried out." In the terms of its entry into the war against Japan on August 8, the U.S.S.R. pledged to support the independence of Korea. On the following day Soviet troops went into action in Manchuria and landed on the northern tip of Korea.

The General Order No. 1, drafted on August 11 by the United States for Japanese surrender terms in Korea, provided for Japanese forces north of latitude 38° N (the 38th parallel) to surrender to the Soviets and those south of that line to the Americans. Stalin did not object to the contents of the order, and on September 8 American troops landed in southern Korea, almost a month after the first Soviet entry. On the following day the United States received the Japanese surrender in Seoul. There were now two zones—northern and southern—for the Soviets had already begun to seal off the 38th parallel.

The southern zone. The end of Japanese rule caused political confusion among Koreans in both zones. In the south various political parties sprang up. Although they were roughly divided into rightists, leftists, and middle-of-the-roaders, they had a common goal: the immediate attainment of self-government. As early as Aug. 16, 1945, some Koreans organized a Committee for the Preparation of Korean Independence, headed by Woon-hyung Lyuh (Yŏ Un-hyŏng), who was closely associated with the leftists. On September 6 the People's Republic of Korea was proclaimed by delegates attending a "national assembly" called by the committee. But the U.S. military government, under Lieutenant General John R. Hodge, the commanding general of the U.S. armed forces in Korea, refused to recognize the republic. The exiled Korean provisional government, on returning, also was compelled to declare itself a political party, not a government. U.S. policy in Korea was to establish a trusteeship that would supersede both the American and the Soviet occupation forces in Korea.

In late December the Council of Foreign Ministers (representing the United States, the Soviet Union, and Great Britain) met in Moscow and decided to create a four-power trusteeship of up to five years. Upon receiving the news, Koreans reacted violently. In February 1946, to soothe the discontent, the military government created the Representative Democratic Council as an advisory body to the military government. This body was composed of Koreans and had as its chairman Syngman Rhee, former president of the Korean government-in-exile.

In October the military government created an Interim Legislative Assembly, half of whose members were elected by the people and half appointed by the military government. The assembly was empowered to enact ordinances on domestic affairs but was subject to the veto of the military government. The feeling against trusteeship came to a climax several months later when the assembly condemned trusteeship in Korea.

The northern zone. Unlike the U.S. forces in the south, the Soviet army marched into the north in 1945 accompanied by a band of expatriate Korean communists.

Soviet policies in the North

By placing the latter in key positions of power, the Soviet Union easily set up a communist-controlled government in the north. On August 25 the People's Executive Committee of South Hamgyŏng province was created by the South Hamgyŏng province Communist Council and other nationalists. The Soviet authorities recognized the committee's administrative power in the province, thus setting a precedent for the committee's role throughout the provinces of the northern zone. In this way the Soviet Union placed the north under its control without actually establishing a military government. In October Korean leaders in the north organized the Bureau of Five Provinces Administration, a central governing body, and this was replaced in February 1946 by the Provisional People's Committee for North Korea. This new agency, a de facto central government, adopted the political structure of the Soviet Union.

Kim Il-sung, who arrived in P'yŏngyang in the uniform of a major of the Red Army, was introduced to the people as a national hero on Oct. 14, 1945. Shortly after his public appearance, Kim was elected first secretary of the North Korean Central Bureau of the Communist Party. After the Provisional People's Committee was organized, with Kim as its chairman, it assumed the helm of existing central administrative bureaus. A year later, in February 1947, a legislative body was established under the name of the Supreme People's Assembly, and, with the strong support of the Soviet occupation authorities, Kim commenced consolidating his political power.

Establishment of the two republics. The Moscow Conference of December 1945, which created the four-power trusteeship, established a Joint U.S.-U.S.S.R. Commission of the rival U.S. and Soviet military commands in Korea for the settlement of the question of a unified Korea. When the commission convened in Seoul from March to May 1946, the Soviet delegates demanded that those Korean political groups that had opposed trusteeship be excluded from consultation. The United States refused, and on this rock foundered all attempts by the commission to prepare for the unification of Korea.

The United States presented the question of Korean unification to the United Nations in September 1947. In November the UN General Assembly adopted a resolution, proposed by the United States, that called for general elections in Korea under the observation of a UN Temporary Commission on Korea. Those elected were to make up a National Assembly, establish a government, and arrange with the occupying powers for the withdrawal of their troops from Korea. The U.S.S.R., however, barred the Temporary Commission from entering the northern zone. The south, however, held elections under the supervision of the Temporary Commission on May 10, 1948. The National Assembly convened on May 31 and elected Rhee as its speaker. Shortly afterward a constitution was adopted, and Rhee was elected president on July 20. Finally, on August 15, the Republic of Korea was inaugurated, with Seoul as the capital, and the military government came to an end. In December the UN General Assembly declared that the republic was the only lawful government in Korea.

Meanwhile, on Nov. 18, 1947, the Supreme People's Assembly of North Korea set up a committee to draft a constitution. The committee adopted the new constitution in April 1948, and on August 25 elections for members of the Supreme People's Assembly were held with a single list of candidates. On September 3 the constitution was ratified by the Supreme People's Assembly, which was holding its first meeting in P'yŏngyang. Kim Il-sung was appointed premier, and on September 9 the Democratic People's Republic of Korea was proclaimed, with the capital at P'yŏngyang. On October 12 the U.S.S.R. recognized this state as the only lawful government in Korea.

(B.-h.H./Y.I.L.)

THE KOREAN WAR

Southern military readiness

South Korea began to organize a police constabulary reserve in 1946. By June 1950, when the war broke out, South Korea had a 98,000-man force equipped only with small arms, which was barely enough to deal with internal revolt and border attacks. The U.S. occupation forces completely withdrew from Korea by June 1949, leaving behind them a force of about 500 men as a U.S. Military Advisory Group to train the South Korean armed forces. In 1949–50 the United States authorized substantial military aid for South Korea, but the military equipment committed under this program was still en route when North Korean troops invaded the South. South Korea was thus unprepared to resist the total invasion from the North.

Military preparations in North Korea had been much more extensive. The Soviet occupation forces left North Korea in December 1948, leaving behind for training purposes 150 advisers for each army division. In March 1949 the U.S.S.R. concluded a reciprocal-aid agreement with North Korea, in which it agreed to furnish heavy military equipment, and by June 1950 North Korean forces numbered 135,000, including a tank brigade. As early as 1946 the Soviets were sending thousands of Koreans to the U.S.S.R. for specialized training, and during 1949–50 China transferred about 12,000 Korean troops from its army to the North Korean forces. The North Korean forces were thus far superior to those of South Korea in training and equipment when, on June 25, 1950, North Korean troops launched a full-scale invasion of the south.

UN intervention. On June 26 (June 25, New York time) the UN Security Council approved a resolution condemning the invasion of South Korea. The Soviet Union was unable to impose a veto, because its delegate had been boycotting the meetings to protest the fact that the People's Republic of China had no representation in the United Nations. On June 27 U.S. President Harry S. Truman issued the order for U.S. air and naval forces to resist communist aggression in Korea; that afternoon the UN Security Council ratified Truman's decision to send air and sea aid to Korea. But Seoul fell on June 28, and most of the South Korean army was destroyed. On June 30 Truman ordered U.S. ground forces in Japan into Korea; the first U.S. troops reached the battlefield on July 4. The UN approved the creation of a unified command in Korea, and General Douglas MacArthur was appointed commander. Sixteen member nations sent armed contingents, but the United States furnished the great bulk of the air units, naval forces, supplies, and financing.

The North Koreans continued to advance, despite the presence of U.S. troops. In early August the UN retreat came to an end in a defense perimeter along the Naktong River, forming a semicircular beachhead around Pusan in Korea's extreme southeast. On September 15 MacArthur counterattacked, catching the communists on the flank by an amphibious attack on Inch'ŏn. North Korean forces were trapped and either surrendered or fled. By October 1 the UN forces were back at the 38th parallel. On September 27 the U.S. Joint Chiefs of Staff had ordered MacArthur to destroy the North Korean armed forces, and two days later Truman authorized him to advance into North Korea. On October 7 the UN General Assembly approved the resolution to permit entry into North Korea. On October 20 the UN forces entered P'yŏngyang, and on October 26 they reached the Chinese border at the Yalu River.

The UN retreat

Chinese intervention. The Chinese, who had moved troops along the Yalu after the Inch'ŏn landing, entered Korea in November in overwhelming numbers. They forced the UN forces to retreat in disorder, and Seoul was reevacuated on Jan. 4, 1951. But the Chinese were halted around P'yŏngt'aek (about 30 miles south of Seoul), and in February the UN General Assembly formally condemned China as an aggressor. The UN counteroffensive began in late January. By March 31 the UN forces had again reached the 38th parallel. MacArthur now publicly advocated an extension of the war to China, but Truman dismissed MacArthur from all of his commands, and General Matthew B. Ridgway took his place. From then until the armistice, the UN forces fought a holding action along the 38th parallel; indeed, in many places the UN forces were slightly north of the line.

Armistice and aid. The Soviet delegate to the United Nations proposed a discussion of a cease-fire and an armistice in June 1951, and in July negotiations began between the United Nations and the communist commanders at Kaesŏng, later resumed at P'anmunjŏm. Many issues stood between the two negotiators. The first was the Chinese demand that all foreign troops be withdrawn from Korea, which was met by a steadfast refusal by the United States. The second issue was the boundary: the communists demanded the restoration of the 38th parallel, but the United States insisted on the existing battle line. The third and most important issue was that of prisoners. The UN forces held 171,000 prisoners, 50,000 of them unwilling to return to their communist countries. The communists were determined to have all prisoners back. On this matter the negotiations were deadlocked and did not resume until after the death of Joseph Stalin in March 1953. A new U.S. administration under President Dwight D. Eisenhower, deeply concerned with balancing the U.S. budget, was determined to end the impasse, even if this involved resuming hostilities. On the other hand, the war-weariness of the communists was increasing. The communists agreed to hand over to a neutral commission the UN-held prisoners of war who did not wish to be repatriated. But Syngman Rhee opposed any agreement that would leave Korea divided and demanded that the military offensive be resumed. In June Rhee suddenly released 27,000 North Korean anticommunist prisoners in defiance of the United Nations, whereupon the communists broke off negotiations. Negotiations were resumed one month later. Rhee acquiesced and agreed to support the armistice even though he would not sign it. In return, the United States promised to extend economic aid and conclude a mutual-security pact to protect South Korea against further aggression.

Armistice signed

The armistice was signed on July 27, 1953. The United Nations had won most of its demands. The military line became the boundary between North and South Korea, and a demilitarized zone was created that extended 1.2 miles (2 kilometres) along each side of the boundary. In addition, commissions were established to enforce the cease-fire regulations. A Neutral Nations Commission for Repatriation was entrusted with the repatriation of prisoners.

The war resulted in roughly 4,000,000 casualties, including civilians. South Korean casualties were some 1,313,000 (1,000,000 civilians); communist casualties were estimated at 2,500,000 (including 1,000,000 civilians). The United States lost about 37,000 in action (the official figure, which had been recorded as some 54,000, was revised in 2000 after it was discovered that a clerk had incorrectly included military noncombatant deaths worldwide), South Korea some 47,000, and the UN forces 3,194; but the estimated losses of China in action were 900,000 men and of North Korea 520,000. During the war, two-fifths of Korea's industrial facilities were destroyed and one-third of its homes devastated. Following the war, the UN Korean Reconstruction Agency (UNKRA) was established to carry out economic aid to South Korea, with most of the contributions being provided by the United States. The UNKRA came to an end in 1958, but UN Emergency Relief and aid from other international voluntary agencies continued.

(B.-h.H./Y.I.L.)

MAP INDEX

NORTH KOREA

Political subdivisions

Chagang . . . . 40 50 N 126 30 E
Hamgyŏng,
North . . . . 42 00 N 129 30 E
Hamgyŏng,
South . . . . 40 00 N 127 30 E
Hwanghae,
North . . . . 38 30 N 126 25 E
Hwanghae,
South . . . . 38 15 N 125 30 E
Hyangsan . . . . . 40 00 N 126 05 E
Kaesŏng . . . . 38 00 N 126 38 E
Kangwŏn . . . . . 38 45 N 127 35 E
Namp'o . . . . 38 45 N 125 25 E
P'yŏngan,
North . . . . 40 00 N 125 15 E
P'yŏngan,
South . . . . 39 30 N 126 24 E
P'yŏngyang . . . . . 39 04 N 125 50 E
Yanggang . . . . . . 41 15 N 128 00 E

Cities and towns

Anak . . . . 38 30 N 125 30 E
Anju . . . . 39 36 N 125 40 E
Aoji . . . . 42 31 N 130 23 E
Chaeryŏng . . . . . 38 24 N 125 37 E
Changjin . . . . 40 23 N 127 15 E
Changyŏn . . . . . . 38 15 N 125 06 E
Chasŏng . . . . 41 28 N 126 38 E
Ch'ŏlsan . . . . 39 46 N 124 40 E
Ch'ŏngjin . . . . . . 41 46 N 129 49 E
Chŏngju . . . . 39 41 N 125 13 E
Chŏngp'yŏng . . . 39 47 N 127 24 E
Ch'osan . . . . 40 50 N 125 48 E
Chunggang
(Chunggangjin) . 41 46 N 126 52 E
Chunghwa . . . . . 38 52 N 125 47 E
Chŭngsan . . . . . . 39 06 N 125 22 E
Haeju . . . . 38 02 N 125 42 E
Hamhŭng . . . . . . 39 54 N 127 32 E
Hoeryŏng . . . . . . 42 26 N 129 45 E
Hoeyang . . . . 38 42 N 127 36 E
Hongwŏn . . . . . . 40 02 N 127 58 E
Huch'ang . . . . . . 41 30 N 127 16 E
Hŭich'ŏn . . . . 40 10 N 126 17 E
Hŭngdŏki
(Hŭngnam) . . . . 39 51 N 127 36 E
Hwadae . . . . 40 49 N 129 30 E
Hwap'yŏng . . . . . 41 15 N 126 53 E
Hyangsan . . . . . . 40 03 N 126 10 E
Hyesan . . . . 41 24 N 128 10 E
Ich'ŏn . . . . 38 29 N 126 53 E
Iwŏn . . . . 40 19 N 128 39 E
Kaesŏng . . . . 37 58 N 126 33 E
Kangdong . . . . . . 39 08 N 126 06 E
Kanggye . . . . 40 58 N 126 36 E
Kangsŏn . . . . 38 54 N 125 34 E
Kapsan . . . . 41 05 N 128 18 E
Kilchu . . . . 40 58 N 129 20 E
Kimch'aek
(Sŏngjin) . . . . . . 40 41 N 129 12 E
Kimhwa . . . . 38 25 N 127 37 E
Kop'ung . . . . 40 36 N 125 57 E
Kosŏng . . . . 38 44 N 128 12 E
Kowŏn . . . . 39 26 N 127 15 E
Kŭmch'ŏn . . . . . . 38 09 N 126 29 E
Kŭmgang . . . . . . 38 37 N 127 59 E
Kusŏng . . . . 39 59 N 125 15 E
Kyŏngsŏng . . . . . 41 35 N 129 37 E
Kyŏngwŏn . . . . . 42 49 N 130 09 E
Maengsan . . . . . . 39 40 N 126 30 E
Mandŏk . . . . 40 36 N 128 34 E
Man'gyŏndae . . . 38 59 N 125 40 E
Manp'o . . . . 41 09 N 126 17 E
Munch'ŏn . . . . . . 39 14 N 127 22 E
Mundŏk . . . . 39 30 N 125 36 E
Musan . . . . 42 14 N 129 13 E
Myŏngch'ŏn . . . . 41 04 N 129 26 E
Myŏnggan . . . . . . 41 15 N 129 30 E
Najin . . . . 42 15 N 130 18 E
Namp'o . . . . 38 44 N 125 24 E
Ongjin . . . . 37 56 N 125 22 E
Onsŏng . . . . 42 57 N 129 59 E
Ŏrang . . . . 41 27 N 129 40 E
Paegam . . . . 41 34 N 128 29 E
Pakch'ŏn . . . . . . 39 44 N 125 35 E
P'anmunjŏm . . . . 37 57 N 126 40 E
Poch'ŏn . . . . 41 31 N 128 18 E
Pukchin . . . . 40 12 N 125 45 E
Pukch'ŏng . . . . . 40 14 N 128 19 E
P'ungsan . . . . . . 40 49 N 128 09 E
Puryŏng . . . . 42 04 N 129 43 E
Puyun . . . . 41 53 N 129 30 E
P'yŏnggang . . . . 38 25 N 127 17 E

P'yŏngsŏng 39 15 N 125 52 E
P'yŏngwŏn 39 18 N 125 36 E
P'yŏngyang 39 01 N 125 45 E
Sariwŏn 38 30 N 125 45 E
Sinch'ang 40 07 N 128 28 E
Sin'gye 38 30 N 126 32 E
Sinhŭng 40 11 N 127 34 E
Sinp'a 41 25 N 127 46 E
Sinp'o 40 02 N 128 12 E
Sinŭiju 40 06 N 124 24 E
Sŏnch'ŏn 39 48 N 124 55 E
Songhwa 38 21 N 125 08 E
Sŏngjin, see Kimch'aek
Songnim 38 44 N 125 38 E
Songwŏn 40 22 N 125 53 E
Sunch'ŏn 39 25 N 125 56 E
Sŭngho 38 59 N 125 58 E
Sup'ung 40 27 N 124 57 E
Taegwan 40 13 N 125 12 E
T'aet'an 38 04 N 125 18 E
Tanch'ŏn 40 28 N 128 55 E
Tŏkch'ŏn 39 45 N 126 18 E
T'ongch'ŏn 38 57 N 127 52 E
T'ŏsan 38 18 N 126 43 E
Ŭiju 40 12 N 124 32 E
Unggi 42 20 N 130 24 E
Unsan 40 06 N 125 54 E
Usi 40 41 N 125 40 E
Wŏnsan 39 10 N 127 26 E
Yodŏk 39 39 N 126 51 E
Yŏmju 39 54 N 124 36 E
Yŏnan 37 54 N 126 10 E
Yonsa 41 54 N 129 00 E

Physical features and points of interest

Amnok, see Yalu
Chaeryŏng, *river* 38 39 N 125 38 E
Changbaek (Changbai) Mountains 41 40 N 128 00 E
Changjin, *river* 41 25 N 127 46 E
Changjin, Lake, *reservoir* 40 30 N 127 12 E
Changsan Cape 38 08 N 124 39 E
Ch'o Island 38 32 N 124 50 E
Ch'ŏngch'ŏn, *river* 39 35 N 125 28 E
East Korea Bay, see Tongjosŏn Bay
Haeju Bay 37 54 N 125 45 E
Hamgyŏng Mountains 41 35 N 129 15 E
Hŏch'ŏn, *river* 41 22 N 128 08 E
Hŭisaek Peak 41 07 N 127 16 E
Imjin, *river* 37 47 N 126 40 E
Japan (East Sea), Sea of 39 00 N 130 30 E
Ka Island 39 33 N 124 40 E
Kaema Highlands 40 40 N 127 00 E
Kangnam Range 41 00 N 126 30 E
Korea Bay 39 00 N 124 00 E
Kŭmgang, Mount 38 39 N 128 07 E
Kwanmo Peak 41 42 N 129 13 E
Mach'ŏl (Paektu) Range 41 30 N 128 30 E
Musu, Point 40 50 N 129 43 E
Myohyang Range 40 02 N 126 17 E
Myŏrak, Mount 38 17 N 126 12 E
Nangnim, Mount 40 21 N 126 55 E
Nangnim Mountains 40 30 N 127 00 E
Ŏnjin, Mount 38 50 N 126 27 E
Ŏrang, Point 41 23 N 129 48 E
Paektu, see Mach'ŏl Range
Paektu (Baitou), Mount 42 00 N 128 03 E
Pidan Island 39 48 N 124 14 E
Pujŏllyŏng Mountains 40 45 N 128 20 E
Pujŏn, *river* 41 07 N 127 31 E
Puksubaek, Mount 40 42 N 127 45 E
Puktae Range 39 22 N 126 40 E
Sinmi Island 39 33 N 124 53 E
Sŏjosŏn (West Korea) Bay 39 15 N 125 00 E
Sup'ung (Shui-feng), Lake, *reservoir* 40 30 N 125 05 E
T'aebaek Mountains 37 40 N 128 50 E
Taedong, *river* 38 42 N 125 15 E
Taedong Bay 38 00 N 125 02 E
Tasa Island 39 47 N 124 25 E
Tongjosŏn (East Korea) Bay 39 30 N 128 00 E
Tumen (Tuman), *river* 42 18 N 130 41 E
West Korea Bay see Sŏjosŏn Bay
Yalu (Amnok), *river* 39 55 N 124 20 E
Yesŏng, *river* 37 53 N 126 24 E
Yŏnghŭng Bay 39 15 N 127 30 E

SOUTH KOREA

Political subdivisions

Cheju 33 25 N 126 30 E
Chŏlla, North 35 45 N 127 15 E
Chŏlla, South 34 45 N 127 00 E
Ch'ungch'ŏng, North 36 45 N 128 00 E
Ch'ungch'ŏng, South 36 30 N 127 00 E
Inch'ŏn 37 28 N 126 43 E
Kangwŏn 37 45 N 128 15 E
Kwangju 35 10 N 126 55 E
Kyŏnggi 37 30 N 127 15 E
Kyŏngsang, North 36 20 N 128 40 E
Kyŏngsang, South 35 15 N 128 30 E
Pusan 35 10 N 129 05 E
Seoul (Sŏul) 37 35 N 127 00 E
Taegu 35 52 N 128 40 E
Taejŏn 36 20 N 127 25 E

Cities and towns

Andong 36 34 N 128 44 E
Anmyŏn 36 32 N 126 22 E
Ansan 37 21 N 126 52 E
Anyang 37 23 N 126 55 E
Chang'an 35 22 N 129 15 E
Changhang 36 01 N 126 42 E
Changhŭng 34 40 N 126 54 E
Changsŏng 35 19 N 126 48 E
Changsŭngp'o 34 52 N 128 44 E
Ch'angwŏn 35 16 N 128 37 E
Chech'ŏn 37 08 N 128 12 E
Cheju 33 31 N 126 32 E
Chido 35 04 N 126 12 E
Chinan 35 47 N 127 26 E
Chinch'ŏn 36 51 N 127 27 E
Chindo 34 29 N 126 16 E
Chinhae 35 03 N 128 40 E
Chinju 35 11 N 128 05 E
Ch'ŏnan 36 48 N 127 09 E
Ch'ŏngdo 35 39 N 128 42 E
Chŏngju 35 34 N 126 51 E
Ch'ŏngju 36 38 N 127 30 E
Chŏngsŏn 37 22 N 128 39 E
Chŏnju 35 49 N 127 09 E
Ch'unch'ŏn 37 52 N 127 44 E
Ch'ungju 36 58 N 127 56 E
Ch'ungmu 34 51 N 128 26 E
Daegu, see Taegu
Hadong 35 04 N 127 45 E
Haenam 34 34 N 126 36 E
Hamyang 35 31 N 127 44 E
Hamyŏl 36 04 N 126 54 E
Hanam 37 32 N 127 12 E
Hoengsŏng 37 29 N 127 59 E
Hongch'ŏn 37 41 N 127 52 E
Ich'ŏn 37 17 N 127 27 E
Inch'ŏn 37 28 N 126 38 E
Inje 38 04 N 128 10 E
Iri 35 56 N 126 57 E
Kamp'o 35 48 N 129 30 E
Kanggyŏng 36 09 N 127 01 E
Kanghwa 37 45 N 126 29 E
Kangnŭng 37 44 N 128 54 E
Kansŏng 38 22 N 128 28 E
Kap'yŏng 37 50 N 127 30 E
Kijang 35 14 N 129 13 E
Kimch'ŏn 36 07 N 128 07 E
Kimhae 35 14 N 128 53 E
Kimje 35 48 N 126 53 E
Kimp'o 37 38 N 126 42 E
Koch'ang 35 26 N 126 42 E
Kŏch'ang 35 41 N 127 55 E
Koesan 36 48 N 127 48 E
Kohŭng 34 37 N 127 17 E
Koksŏng 35 17 N 127 18 E
Kongju 36 27 N 127 08 E
Koryŏng 35 44 N 128 16 E
Kŭmch'on 37 46 N 126 46 E
Kumi 36 08 N 128 20 E
Kŭmsan 36 06 N 127 29 E
Kunp'o 37 22 N 126 57 E
Kunsan 35 59 N 126 43 E
Kunwi 36 14 N 128 33 E
Kuri 37 35 N 127 09 E
Kwach'ŏn 37 26 N 127 00 E
Kwangju 35 09 N 126 55 E
Kwangmyŏng 37 28 N 126 45 E
Kyŏngju 35 50 N 129 13 E
Kyŏngsan 35 49 N 128 44 E
Masan 35 11 N 128 34 E
Migŭm 37 37 N 127 11 E
Miryang 35 29 N 128 45 E
Mokp'o 34 47 N 126 23 E
Muan 34 59 N 126 29 E
Muju 36 00 N 127 40 E
Mun'gyŏng 36 44 N 128 07 E
Munsan 37 51 N 126 47 E
Naju 35 02 N 126 43 E
Namhae 34 50 N 127 54 E
Namwŏn 35 24 N 127 23 E
Nonsan 36 12 N 127 05 E
Onyang 36 46 N 127 01 E
Osan 37 09 N 127 04 E
P'aju 37 50 N 126 49 E
P'anmunjŏm 37 57 N 126 40 E
P'ohang 36 02 N 129 22 E
Poŭn 36 29 N 127 43 E
Puan 35 44 N 126 44 E
Puch'ŏn 37 29 N 126 47 E
Pusan 35 06 N 129 03 E
Puyŏ 36 17 N 126 54 E
P'yŏnghae 36 43 N 129 27 E
P'yŏngtaek 36 59 N 127 05 E
Samch'ŏk 37 27 N 129 10 E
Samch'ŏnp'o 34 55 N 128 04 E
Sangju 36 25 N 128 10 E
Seoul (Sŏul) 37 34 N 127 00 E
Sŏch'ŏn 36 04 N 126 41 E
Sŏgwip'o 33 15 N 126 34 E
Sokch'o 38 12 N 128 36 E
Sŏngju 35 55 N 128 17 E
Sŏngnam 37 26 N 127 08 E
Songt'an 37 03 N 127 04 E
Sŏnsan 36 14 N 128 18 E
Sŏsan 36 47 N 126 27 E
Sunch'ŏn 34 57 N 127 29 E
Suwŏn 37 16 N 127 01 E
T'aean 36 45 N 126 18 E
T'aebaek 37 10 N 128 59 E
Taech'ŏn 36 21 N 126 36 E
Taegu (Daegu or Taiku) 35 52 N 128 36 E
Taejŏn 36 20 N 127 26 E
Tangjin 36 54 N 126 38 E
Tongduch'ŏn 37 55 N 127 03 E
Tonghae 37 32 N 129 06 E
Tonggwangyang 34 56 N 127 44 E
Ŭijŏngbu 37 44 N 127 02 E
Ŭiryŏng 35 19 N 128 16 E
Ŭisŏng 36 21 N 128 42 E
Ulchin 36 59 N 129 24 E
Ulsan 35 33 N 129 19 E
Wando 34 18 N 126 45 E
Wŏnju 37 21 N 127 58 E
Yanggu 38 06 N 127 59 E
Yangyang 38 04 N 128 38 E
Yech'ŏn 36 39 N 128 27 E
Yesan 36 41 N 126 51 E
Yŏch'ŏn 34 45 N 127 40 E
Yŏju 37 18 N 127 38 E
Yŏnch'ŏn 38 06 N 127 04 E
Yŏngch'ŏn 35 58 N 128 56 E
Yŏngdŏk 36 24 N 129 22 E
Yŏnggwang 35 16 N 126 31 E
Yŏngju 36 49 N 128 37 E
Yŏngwŏl 37 11 N 128 28 E
Yŏngyang 36 40 N 129 07 E
Yŏsu 34 44 N 127 44 E

Physical features and points of interest

Andong Lake, *reservoir* 36 40 N 128 52 E
Anmyŏn Island 36 30 N 126 22 E
Asan Bay 36 56 N 126 51 E
Changgi, Cape 36 05 N 129 34 E
Cheju (Quelpart) Island 33 20 N 126 30 E
Cheju Strait 33 40 N 126 28 E
Chin Island 34 25 N 126 15 E
Ch'o Island 34 14 N 127 15 E
Haenam Peninsula 34 40 N 126 25 E
Hajo Island 34 17 N 126 03 E
Halla, Mount 33 22 N 126 32 E
Hallyŏ Marine National Park 34 45 N 128 30 E
Han, *river* 37 45 N 126 11 E
Imja Island 35 05 N 126 05 E
Imjin, *river* 37 47 N 126 40 E
Japan (East Sea), Sea of 39 00 N 130 30 E
Kanghwa Island 37 40 N 126 27 E
Kohŭng Peninsula 34 35 N 127 20 E
Kŏje Island 34 52 N 128 37 E
Kŏmun Island 34 18 N 127 24 E
Korea Strait 34 30 N 129 00 E
Kum, *river* 36 00 N 126 40 E
Kŭmŏ Island 34 31 N 127 45 E
Kyŏnggi Bay 37 25 N 126 00 E
Mount Chiri National Park 35 18 N 127 37 E
Mount Halla National Park 33 18 N 126 32 E
Mount Odae National Park 37 45 N 128 37 E
Mount Songni National Park 36 42 N 127 57 E
Mount Sŏrak National Park 38 07 N 128 25 E
Mount Worak National Park 36 51 N 128 10 E
Naktong, *river* 35 07 N 128 57 E
Namhae Island 34 48 N 127 57 E
Paengnyŏng Island 37 57 N 124 40 E
Pukhan, *river* 37 31 N 127 18 E
Quelpart, see Cheju Island
Soan Islands 34 10 N 126 35 E
Sobaek Mountains 36 00 N 128 00 E
Sŏmjin, *river* 34 58 N 127 46 E
Sŏsan, see T'aean Marine National Park
Tadohae Marine National Park 34 30 N 126 40 E
T'aean (Sŏsan) Marine National Park 36 43 N 126 15 E
T'aebaek, Mount 37 06 N 128 55 E
T'aebaek Mountains 37 40 N 128 50 E
Taech'ŏng Island 37 49 N 124 43 E
Taehŭksan Islands 34 41 N 125 26 E
Tolsan Island 34 38 N 127 45 E
Tŏkchŏk Islands 37 14 N 126 07 E
Tsushima Strait 34 00 N 130 00 E
Ullŭng Island 37 29 N 130 52 E
Yellow Sea 36 00 N 124 00 E
Yŏngil Bay 36 02 N 129 26 E
Yŏngsan, *river* 34 54 N 126 32 E

Cities over 5,000,000
Cities 1,000,000 to 5,000,000
Cities 250,000 to 1,000,000
Cities under 250,000
Other localities
National capitals
Political subdivision capitals
PUSAN Political subdivision names
International boundaries
Political subdivision boundaries
Dikes
Dams
National parks
Spot elevations in metres (1 m = 3.28 ft)
The cities of Taejŏn, Kwangju, and Taegu are also capitals of South Ch'ungch'ŏng, South Chŏlla, and North Kyŏngsang, respectively.
Scale 1:4,273,000
1 inch equals approx. 67 miles
Conic Projection
CHINA
RUSSIA
NORTH KOREA
SOUTH KOREA
JAPAN
EAST SEA (SEA OF JAPAN)
YELLOW SEA
Korea Bay
Korea Strait
Tsushima Strait
Cheju Strait
Demilitarized Zone is 1.2 miles (2 kilometres) wide on either side of the military demarcation line.
Seoul
P'yŏngyang
Pusan
Taegu
Inch'ŏn
Kwangju
Taejŏn
Ulsan
Wŏnsan
Hamhŭng
Ch'ŏngjin
Kaesŏng
Yanji
T'ung-hua
KANGNAM RANGE
NANGNIM MOUNTAINS
KAEMA HIGHLANDS
HAMGYŎNG MOUNTAINS
MACH'ŎL RANGE
CHANGBAEK MTS.
Mount Paektu 2750
T'AEBAEK MOUNTAINS
SOBAEK MOUNTAINS
MYOHYANG RANGE
PUJŎLLYŎNG MOUNTAINS
NORTH HAMGYŎNG
SOUTH HAMGYŎNG
YANGGANG
CHAGANG
NORTH P'YŎNGAN
SOUTH P'YŎNGAN
NORTH HWANGHAE
SOUTH HWANGHAE
KANGWŎN
KYŎNGGI
NORTH CH'UNGCH'ŎNG
SOUTH CH'UNGCH'ŎNG
NORTH KYŎNGSANG
SOUTH KYŎNGSANG
NORTH CHŎLLA
SOUTH CHŎLLA
CHEJU
Cheju
Mount Halla 1950
ULLŬNG ISLAND
TSUSHIMA (Japan)
© 2005 Encyclopædia Britannica, Inc.

THE MODERN STATES OF KOREA

North Korea

The Democratic People's Republic of Korea (Korean: Chosŏn Minjujuŭi In'min Konghwaguk), commonly called North Korea, occupies the northern portion of the Korean peninsula, which juts out from the Asian mainland between the East Sea (Sea of Japan) and the Yellow Sea. The country is bordered by China and Russia to the north and by the Republic of Korea (South Korea) to the south. North Korea has an area of 47,399 square miles (122,762 square kilometres), occupying about 55 percent of the peninsula. The national capital, P'yŏngyang, is a major industrial and transport centre near the west coast.

PHYSICAL AND HUMAN GEOGRAPHY

Mountains and valleys

The land. *Relief.* Mountains and valleys characterize most of the country. The Kaema Highlands in the northeast have an average elevation of 3,300 feet (1,000 metres) above sea level and form the topographic roof of the entire Korean peninsula. Mount Paektu (9,022 feet [2,750 metres]), the highest mountain in North Korea and on the peninsula, rises at the northern edge of this plateau; it is an extinct volcano topped by a large crater lake. The Nangnim Mountains run from north to south through the middle of the country, forming a divide between the eastern and western slopes of the peninsula. The Kangnam and Myohyang ranges and Mounts Ŏnjin and Myŏrak, all structural extensions of the Nangnim Mountains, extend parallel to each other toward the southwest. Large river-valley plains have developed between the western mountains; they merge along the narrow, irregular coastal plain on the west coast. The Hamgyŏng Mountains, extending from the Nangnim Mountains to the northeast, form a steep slope between the Kaema Highlands and the East Sea. The T'aebaek Mountains extend from southeastern North Korea into South Korea along the eastern coast.

Drainage and soils. The longest river of North Korea is the Yalu, in Korean called the Amnok. It rises on the southern slope of Mount Paektu and flows southwestward for some 500 miles (800 kilometres) to its mouth on Korea Bay. The Tumen (Tuman) River also begins at Mount Paektu but runs northeastward for 324 miles to the Sea of Japan. There are no large streams along the east coast except for the Tumen River, and all the significant rivers, such as the Yalu, Ch'ŏngch'ŏn, Taedong, Chaeryŏng, and Yesŏng, drain to the Yellow Sea. The relatively large valley plains of the western rivers are major agricultural regions.

More than three-fifths of the soils are locally derived from the weathering of granitic rocks or various kinds of schists (crystalline rocks). The soils are generally brownish, abundant in sandy materials, and low in fertility. Well-developed reddish brown soils derived from limestone are found in North Hwanghae province and the southern part of South P'yŏngan province. Podzols (ash-gray forest soil) have developed in the Kaema Highlands as a result of the cold climate and coniferous forest cover there. Although most of the soils are infertile and lack organic content, the valleys and coastal plains have relatively rich alluvial soils.

Climate. North Korea has a generally cool continental climate. The winter season, from December to March, is long and cold; mean temperatures in January range between 21° F (−6° C) in the south and −8° F (−22° C) in the northern interior. The summer, from June to September, is warm, with mean July temperatures above 68° F (20° C) in most places. Accordingly, the annual range of temperatures is large—about 54° F (30° C) at P'yŏngyang and about 77° F (43° C) at Chunggang (Chunggangjin), where the lowest temperature in the Korean peninsula, −46.5° F (−43.6° C), has been recorded. Winter temperatures on the east coast are some 5° to 7° F (3° to 4° C) higher than those of the west coast.

Precipitation

Most of the country receives about 40 inches (1,000 millimetres) of precipitation annually. The northern inland plateau, however, receives only some 24 inches and the lower reaches of the Taedong River valley 32 inches, while the upper Ch'ŏngch'ŏn River area averages between 48 and 52 inches yearly. Some three-fifths of the annual precipitation falls in the four months from June to September; this heavy concentration of rainfall is related to the humid summer monsoon from the Pacific, which also produces occasional typhoons. Only a small portion of the total precipitation occurs in winter, generally as snow; snowfall can be locally heavy, as in the T'aebaek Mountains.

Plant and animal life. Vegetation on the highlands, especially around Mount Paektu, consists of coniferous trees, such as the Siberian fir, spruce, pine, and Korean pine (Pinus koraiensis). The western lowlands were originally covered by temperate mixed forests with many types of plants, but continuous deforestation has left only remote patches of the original forests. Most of the lowlands are now cultivated, except for some of the hills that are covered with small pine groves mixed with oaks, lindens, maples, and birches. Along streams that are subject to flooding or where the ground is too stony for cultivation, reeds, sedges, wild mulberry trees, and Italian poplars are found. Common river fish include carp and eels.

Because of deforestation, the populations of deer, mountain antelope, goats, tigers, and leopards have greatly decreased and are restricted to the remote forests. In the plains, however, it is still possible to see wild pigeons, herons, cranes (which nest near human habitations), and many migratory waterfowl, which alight in the rice fields.

Settlement patterns. Close examination reveals numerous distinct settlement regions, each with a different natural environment and historical background. Of the eight Korean provinces of the Chosŏn (Yi) dynasty (1392–1910), North Korea contains the three provinces of P'yŏngan, Hwanghae, and Hamgyŏng and the northern parts

© Hiroji Kubota/Magnum Photos

Crater lake at the summit of Mount Paektu, northern Yanggang province, North Korea.

of Kangwŏn and Kyŏnggi provinces. Each province not only was a political unit but also had characteristics of a cultural region in terms of dialect, customs, and way of life. North Korea may also be divided into the two larger traditional regions: Kwansŏ to the west and Kwanbuk to the east, roughly divided by the Nangnim Mountains. Kwansŏ comprises the current provinces of North P'yŏngan, South P'yŏngan, North Hwanghae, South Hwanghae, and Chagang, while Kwanbuk includes North Hamgyŏng, South Hamgyŏng, Yanggang, and Kangwŏn provinces.

Most of the rural population inhabits the eastern and western coastal lowlands and river-valley plains. The inland areas of Chagang and Yanggang provinces are sparsely settled because of the lack of arable land and the cold climate, which does not permit the cultivation of rice. Villages in the lowlands and valley plains usually are clustered together at the bases of the southern slopes of hills, which offer protection against the cold northwesterly winter wind. Scattered fields are tilled by a small number of shifting cultivators in the Kaema Highlands, especially in Yanggang province. The upper Yalu and Tumen river valleys contain settlements associated with lumbering, and fishing villages are numerous along the coasts.

Urban settlement

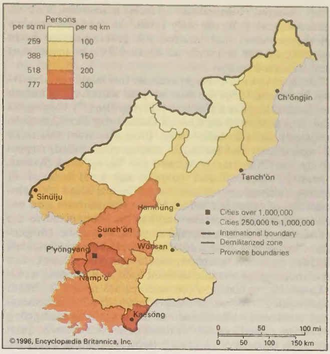
Cities that developed during the Japanese occupation (1910–45) were largely associated with the exploitation of natural resources, industry, and transportation. The communist regime's heavy emphasis on manufacturing resulted in the continuous expansion of the early industrial centres and caused a population flow into the urban areas from the countryside. Most of the cities were destroyed during the Korean War (1950–53) and have since been rebuilt. Urbanization increased rapidly after the war, especially in the period 1953–60. P'yŏngyang is by far North Korea's largest city, being some three times larger than Hamhŭng, the country's second largest city.

The people. *Ethnic and linguistic composition.* The Korean peninsula is one of the most ethnically homogeneous regions in the world. The North Korean population, which has been largely isolated since 1945, is almost entirely Korean; a tiny number of Chinese constitute the only other significant ethnic group. All Koreans speak the Korean language, which is related to Japanese and contains Chinese loanwords. The Korean script, known in North Korea as Chosŏn muntcha and in South Korea as Hangul (*han'gŭl*), is composed of phonetic symbols for the 10 vowels and 14 consonants. In North Korea a systematic effort has been made to eliminate Chinese and Western loanwords, as well as any vestiges of the Japanese imposed during the colonial period, and Chosŏn muntcha has been used exclusively without Chinese characters in newspapers and other publications since 1945.

Religion. The way of life and the value system of Koreans are based fundamentally on Confucian thought. To a lesser extent, Buddhism is also important. Roman Catholic and Protestant missionaries arrived in the 18th and 19th centuries, respectively. The Japanese occupation brought increasing repression of Christianity, and by the end of World War II there were no foreign missionaries left in the country.

The monotheistic religion of Ch'ŏndogyo ("Teaching of the Heavenly Way"), originally known as Tonghak ("Eastern Learning"), was founded by the Confucian teacher Ch'oe Che-u in 1860. A combination largely of Buddhism, Confucianism, and Christianity, Ch'ŏndogyo played a leading role in the March 1st Movement of 1919. Shamanism—the religious belief in gods, demons, and ancestral spirits responsive to a priest, or shaman—existed in Korea before the introduction of Buddhism and Confucianism; its practice has nearly died out in North Korea.

The communist regime has constitutionally confirmed freedom of religion but does not practice it. Ch'ŏndogyo, however, has been used for propaganda purposes, and since the late 1980s Christianity has been used as a means for contact with South Korea and the West. After the Korean War, churches and Buddhist temples were confiscated and looted, and many were converted to other purposes. Religious activity generally has remained under state control, although since the late 1980s there has been some increase in independent practices.



Population density of North Korea.

Demographic trends. North Korea's population grew fairly rapidly after the Korean War, roughly doubling in size between 1953 and 1980. Although the rate of population increase has slowed somewhat since 1970, it is nearly twice that of South Korea. North Korea, however, is underpopulated by East Asian standards, with an overall density of only about two-fifths that in the south.

Population distribution

The population of North Korea is unevenly distributed, with heavy concentrations along the coastlines and only sparse settlement in the interior. This imbalance has been exacerbated by the government's emphasis on industrialization since 1945, which has promoted migration to the cities and, conversely, has produced a severe farm labour shortage. Large numbers of Koreans emigrated from the peninsula during the first half of the 20th century, mainly to China, Siberia, Japan, and the United States. In the 1960s and '70s North Korea conducted a campaign to repatriate Koreans living in Japan; more than 90,000 people took part in the program.

The economy. North Korea has a command (centralized) economy. The means of production are controlled by the state, and priorities and emphases in economic development are set by the government. Since 1954, economic policy has been promulgated through a series of national economic plans. The early plans gave high priority to reconstruction and the development of heavy industries, especially chemicals and metals. Subsequent plans focused on resource exploitation and improving technology, mechanization, and infrastructure. Little attention was given to agriculture until the 1970s, and it was not until the late 1980s that much effort was made to improve the quality and quantity of consumer goods.

Reliable information on the performance of the North Korean economy usually has been lacking. Outside observers have concluded, however, that the country has consistently failed to meet its stated goals and that production statistics quoted by the government often have been inflated. The economy grew rapidly in the first decade after the Korean War but then tended to stagnate or to grow only slowly; by the early 1990s North Korea was experiencing years of economic decline, in large part the result of the demise of the Soviet Union and the communist nations of Europe that had been the country's largest trading partners.

Overall, North Korea has changed from an agricultural to an industrial nation since World War II. Economic goals have been linked to the general government policy of self-reliance (*juche*, or *chuch'e*). The country has shunned foreign investment, although it has accepted considerable economic aid from the former Soviet Union, China, and eastern European countries. Despite its stated policy of self-reliance, North Korea has found it necessary to import such essential commodities as fuels, machinery, and, on occasion, grains. By the early 1990s, the poor performance of the economy had induced the government to begin opening up the economy to limited foreign investment and increased trade.

Efforts have been made to increase low labour productivity. In the late 1950s the state adopted a mass-mobilization measure called the Ch'ŏllima ("Flying Horse") Movement that was patterned on China's Great Leap Forward. Subsequently, in the early 1960s programs were instituted in agricultural and industrial management called, respectively, the Ch'ongsan-ni Method and Taean Work System.

Mining. North Korea contains the great bulk of all known mineral deposits on the peninsula; some 200 minerals are of economic value. Most important are iron ore and coal, although greater emphasis is being given to the extraction of such minerals as gold, magnesite (magnesium carbonate), lead, and zinc. Other abundant minerals include tungsten, graphite, barite (barium sulfate), and molybdenum (a metallic element used in hardening steel).

Iron and coal

Large, high-grade iron ore reserves are mined in North and South Hwanghae, South P'yŏngan, and South Hamgyŏng provinces, while deposits at Musan, North Hamgyŏng province, mined extensively for decades, are of lower quality. Rich deposits of anthracite (hard coal) occur along the Taedong River—notably at Anju, north of P'yŏngyang—and near Paegam in Yanggang province. There also is some lignite (brown coal) in the far northeast and at Anju. North Korea's magnesite deposits, the largest in the world, are centred on Tanch'ŏn, in South Hamgyŏng province.

Forestry and fishing. The northern interior contains large forest reserves of larch, spruce, and pine trees. Most of the coastal slopes have been extensively deforested, however, much of this done by the Japanese during World War II; reforestation programs have stressed economic forestry. Forestry production, having declined after the war, has not grown substantially.

The sea is the main source of protein for North Koreans, and the government has continually expanded commercial fishing since the 1950s. Most fishing activity centres on the coastal areas on each side of the peninsula, although deep-sea fishing has expanded since the 1970s. The main species caught include pollack, sardines, mackerel, herring, pike, yellowtail, and shellfish. Aquaculture has increased in importance.

Agriculture. By 1958 all privately owned farms were incorporated into more than 3,000 cooperatives; each cooperative comprises about 300 families on about 1,200 acres (500 hectares). The farm units are controlled by management committees, which issue orders to the work teams, set the type and amount of seed and fertilizer to be used, and establish production quotas. Produce is delivered to the government, which controls distribution through state stores.

Despite the decreasing proportion that agriculture contributes to the national economy (about one-fourth in 1990), there has been an increase in cultivated land, irrigation projects, the use of chemical fertilizers, and mechanization. Farmers are paid for their labour in money or in kind and are allowed to keep chickens, bees, fruit trees, and gardens. Surplus produce can be sold at local markets held periodically.

The main food crops are grains—notably rice, corn (maize), millet, wheat, and barley. Grain production has increased considerably since the 1950s, and the country is self-sufficient in rice; wheat, however, must still be imported. Sweet potatoes, soybeans, vegetables, and tree fruits are raised extensively. Industrial crops include tobacco, cotton, flax, and rape (an herb grown for its oilseeds). Livestock raising is concentrated in areas poorly suited for crop raising. Livestock production has increased steadily, especially poultry production.

Livestock

Industry and power. The industrial sector is organized into state-owned enterprises and production cooperatives, the latter being confined largely to handicrafts, marine processing, and other small-scale operations. The most important industries are iron and steel, machinery, chemicals, and textiles. Iron and steel production initially was centred at Songnim and Ch'ŏngjin but has been expanded to include the large integrated mill at Kimch'aek. Industrial and agricultural machinery is manufactured at Kangsŏn, near P'yŏngyang, and several other cities, including Hŭich'ŏn. The production of chemicals is focused on fertilizers and petrochemicals, much of the latter being manufactured in the Anju area north of P'yŏngyang. The textile industry is centred at P'yŏngyang, Sinŭiju, and Sunch'ŏn. Other products include armaments, vehicles, cement, glass, ceramics, and some consumer goods (mainly clothing and processed food).

Industrial development is related to the country's large supply of electric power. Hydroelectric power resources were heavily developed during the Japanese regime along the Yalu River and its upper tributaries. Power production is still based mainly on hydroelectricity, but thermal electricity is becoming important because of lower construction costs and the unreliability of hydroelectric power during the dry season. Production of electricity has increased steadily, although it has not kept pace with industrial growth; nonetheless, electricity is sold to China.

Finance and trade. The Central Bank of the Democratic People's Republic of Korea is the sole bank of issue. It receives all national revenues and precious metals and provides government agencies with working capital. There are several other state banks, all supervised by the Central Bank. A limited amount of joint-venture banking has been allowed since the late 1980s.

Foreign trade has expanded and diversified slowly. At first trade was conducted only with the Soviet Union and China, but since the 1960s it has been allowed with a growing number of countries. Since the demise of the Soviet Union, trade with Russia, though declining, has remained important. In addition to China, other major trading partners include Japan, Germany, and Hong Kong. Trade also has been conducted with South Korea since 1990. Imports mainly consist of fuels, machinery (including machine tools and transport equipment), chemicals, textile yarns and fabrics, and grains. Exports include minerals (*e.g.,* lead, magnesite, and zinc), metallurgical products (iron and steel, nonmetallic minerals), and cement.

Transportation. Railways are the principal means of transportation. The basic railway pattern runs in a north-south direction, roughly parallel to the coasts, with branch lines to the river valleys. Because of the high mountains, there is only one east-west railway line, between P'yŏngyang and Wŏnsan. The west-coast line runs from Kaesŏng near the South Korean border to Sinŭiju on the Chinese border, connecting the major cities. From this major line a branch from P'yŏngyang southwestward to Namp'o connects centres of machine building and foundries. Another line runs northward from P'yŏngyang to Manp'o on the Yalu River, connecting the western interior to China's northeastern provinces. The major railway on the east coast runs from Wŏnsan northward to Najin and continues to Namyang on the Chinese border.

Railways

Highway transportation is not as important as railroads, because few motor vehicles are available. Major roads parallel the rail lines. Express highways connect P'yŏngyang with Wŏnsan, Namp'o, and Kaesŏng. Most roads, however, are not paved.

River transportation plays an important role in transporting agricultural products, minerals, and passengers. The most important rivers utilized for freight transportation are the Yalu, Taedong, and Chaeryŏng. Namp'o—the entry port to P'yŏngyang—Haeju, and Tasa are the major ports on the west coast, as are Wŏnsan, Hŭngnam (outport of Hamhŭng), Ch'ŏngjin, and Najin in the east.

Air services are controlled by the air force. Flights are maintained between the major cities, and international services connect P'yŏngyang with Peking (Beijing) and Moscow. Sunan Airport, northwest of P'yŏngyang, serves as the international airport.

Administration and social conditions. *Government.* The first constitution of the Democratic People's Republic of Korea was promulgated in 1948 and was replaced with a new constitution in 1972; in 1992 revisions reportedly were made to the 1972 document, but details of those changes have not officially been made known. The Supreme People's Assembly, constitutionally the highest organ of state power, consists of a single chamber with 687 members; it also has a 15-member Standing Committee that meets when the assembly is not in session. Executive authority rests with a president (who is head of state), one or more vice presidents, and a 45-member cabinet called the Central People's Committee. Actual authority, however, resides in the extraconstitutional political body of the Korean Workers' Party (KWP). The government is highly centralized and totalitarian in nature.

The judicial system consists of the courts and a procuracy. Judicial authority rests with the Central Court, whose judges are elected for three-year terms by the Supreme People's Assembly, and with a number of lesser provincial and people's courts, whose members are elected by local people's assemblies. Judges usually are party members or are controlled by the party. The Central Procurator's Office prosecutes cases and maintains surveillance over all citizens; it is headed by a procurator-general, who is selected by the Supreme People's Assembly. The courts and procuracy are independent of each other, and actually the procurator's office functions as the fourth branch of government.

Local government

North Korea is divided administratively into nine provinces (*do* or *to*), the three special province-level cities (*chikalsi*, or *jikalsi*) of Kaesŏng, Namp'o, and P'yŏngyang, and one special region (*chigu*) of Hyangsan. These units are further subdivided into cities (*si*), counties (*gun*, or *kun*), and villages (*ri*), the smallest administrative unit. Local people's assemblies elect the members of their people's committees, which make local economic plans and budgets with the approval of higher authorities.

There are a number of political parties and social organizations that serve to support the KWP. All political activities are sponsored by the party or require its sanction and must closely follow the party line and policies. Elections provide a means whereby assent is registered for the policy and program of the party; they do not allow freedom of expression. There is seldom more than one candidate on the ballot for each constituency, and the electoral system is completely controlled by the party.

Armed forces and security. Since 1966 there has been an emphasis on military preparedness, and economic plans have been altered to support high military expenditures. North Korea has one of the world's largest military organizations. The army is by far the largest force; there are also an air force and a navy. All men and a limited number of women are subject to conscription; the duration of service officially is three to four years but typically lasts longer. There also are large reserve and paramilitary forces.

The country's internal-security system is large and extensive. People's movements and social activities are monitored and controlled even down to the household level. The Ministry of Public Security functions as a national constabulary, while political control and counterintelligence are the responsibility of the State Security Department.

Education. Education serves as a process of indoctrination in communist ideology and a means to supply skilled workers, technicians, and scientists to meet the government's economic goals. All students are required to engage in productive labour along with their studies, which emphasize science and technology. In 1967 education was made compulsory for those between the ages of 7 and 16, later changed to between 5 and 16. The system comprises one year of preschool, four years of primary school, and six years of secondary school. Institutions of higher education offer programs of two to six years in length; the most important school is Kim Il-sung University in P'yŏngyang. There is also a well-developed system of adult education, the major components of which are technical schools located in large industrial centres.

Health and welfare. Medical care is free, but there is a shortage of physicians and medicine. Medical benefits are provided by social insurance for workers who are temporarily or permanently disabled and women during pregnancy and childbirth. There are also funeral benefits and old-age pensions. Homes for the aged in each province have been in operation since 1964.

Housing

Reconstruction of houses after the Korean War was given high priority, and dwellings have improved considerably. Rural mud-walled, thatched-roofed huts have been replaced by brick buildings with tile or slate roofs. Urban housing is classified into five groups that range from one room and a half-sized kitchen to free-standing houses with gardens. Workers are expected to live in apartments rather than houses, and housing projects are supported almost solely by the government. Heating systems in the apartments and urban water supplies are inadequate. Numerous high-rise buildings have been constructed in the larger cities since the 1970s, especially in P'yŏngyang. City streets are strangely empty of pedestrians, as the North Koreans have few leisure hours.

Because of the high priority of industrialization and defense, the provision of consumer goods and social services has long been inadequate. The material economy and the lot of the peasant have improved since World War II, however, with adequate supplies of such basic goods as food and clothing generally being available. North Korea was established as a classless society, but a new, privileged elite of high-ranking KWP officials has emerged.

Cultural life. The compound religious strains of shamanism, Buddhism, Confucianism, and Taoism have deep roots in Korean culture. Although the country has received continuous streams of foreign cultural influence mainly from China, Koreans have kept their identity and maintained and developed their unique language and customs. Westernization, begun in the late 19th century, was accomplished in harmony with Korean tradition and slowly transformed the culture without much conflict until the 1940s.

After World War II the occupying Soviets did not recognize the Korean traditional family system or Confucian philosophy; age-old lineage records were burned, and the kinship system was broken. Through education, people were molded to fit the pattern of party idealism, and private life and individual freedom became extremely limited. Development plans since the Korean War have demanded almost superhuman patience and labour from the North Koreans. As a result, the people have had to lead an austere existence. The standard of living has improved, but leisure and cultural activities have continued to be regimented and geared toward organized group activities, such as rallies and museum tours.

The arts. The government is heavily involved with maintaining and advancing the traditional fine arts and other cultural features as an expression of nationalism. The selection of cultural items is based on communist ideology, and writers and artists attempt to propagate the superiority and independence of Korean culture. All North Korean writers, artists, dancers, and musicians are assigned to government institutions such as the National Theatre for the Arts, National Orchestra, and National Dancing Theatre in P'yŏngyang and provincial organizations of music, ballet, and drama. Museums have been well-supported by the government, and many archaeological sites have been excavated to promote the growth of a strong nationalistic feeling. Among the country's most notable museums are the Korean Revolutionary Museum and the Korean Fine Arts Museum in the capital. Archaeological sites include those located in the Nangnang district of P'yŏngyang and at Kungsan, near Yonggang.

Museums

The press and broadcasting. Of the daily newspapers, the *Rodong* (or *Nodong*) *sinmun* ("Labour News"), published by the KWP Central Committee, and the government's *Minju Chosŏn* ("Democratic Korea") have the largest circulations. The monthly *Kŭlloja* ("Workers") of the KWP Central Committee is one of the most influential periodicals. The Korean Central News Agency controls the dissemination of information, and all papers are strictly censored. The government long has recognized the importance of radio and television as mass media, and they have played a great role in ideological education. Radio broadcasts reach all parts of the country. Almost all North Korean households have access to radio broadcasts as a result of a government project to link household loudspeakers to village receivers. Television broadcasting in North Korea also has been made available to all parts of the country, and the number of television sets has increased.

(C.Le./W.-i.Y.)

For statistical data on the land and people of North Korea, see the *Britannica World Data* section in the BRITANNICA BOOK OF THE YEAR.

HISTORY

In 1948, when the Democratic People's Republic of Korea was established, Kim Il-sung became the first premier of the North Korean communist regime. In 1949 he became chairman of the KWP, created from communist parties founded earlier. Until his death in July 1994, Kim ruled by promoting a personality cult centred on himself as the "Great Leader" of the Korean people.

The 1950s and '60s. In the aftermath of the Korean War, Kim purged the so-called domestic faction—an indigenous communist group that had survived the colonial period. After 1956, as the Sino-Soviet conflict intensified, Kim shifted his positions vis-à-vis Moscow and Peking three times: from pro-Soviet to neutral, to pro-Chinese, and finally to independent. During 1956–58, he purged the pro-Chinese Yenan faction and eliminated a pro-Soviet faction from the KWP Central Committee.

In 1966, after a visit to P'yŏngyang by the Soviet prime minister, Aleksey N. Kosygin, Kim announced what became known as the independent party line in North Korea, which stressed the principles of "complete equality, sovereignty, mutual respect, and noninterference among the communist and workers' parties." From this party line, KWP theoreticians evolved four self-reliance (*juche*) principles: "autonomy in ideology, independence in politics, self-sufficiency in economy, and self-reliance in defense."

The four principles

In the late 1960s the regime implemented a program for strengthening the armed forces. More military airfields were constructed and large underground aircraft hangars were built. In addition, a large standing army and a strong militia were maintained. North Korea's emphasis on its military forces proceeded hand in hand with its continued stress on developing a self-reliant economy. With aid from the Soviet Union, China, and eastern Europe, North Korea implemented a series of economic development plans. But as aid declined sharply—first from the Soviet Union beginning in the late 1950s and then from China at the start of the Cultural Revolution—the seven-year plan of 1961–67 was seriously affected, as indicated by its extension for three years. (B.-h.H./Y.I.L.)

North Korea since 1970. *Domestic developments.* Subsequent economic plans also failed to achieve their stated goals. While the country's economic growth was hampered by the decline in foreign aid and its heavy expenditures on defense, the continued priority assigned to heavy industry created a severe shortage of daily commodities and lowered living standards. Food shortages were aggravated, in part, because of a more than threefold increase in population since the 1950s. Famine ravaged the country, especially in the mid-1990s, although its severity was mitigated by aid from the UN and various countries. During the same period, resources were put into developing long-range missiles.

When the 1972 constitution was adopted, the premiership was changed to a presidency, which Kim Il-sung assumed; Kim also was retained as the chairman (renamed the secretary-general) of the KWP. In 1980 the KWP held its first party congress in a decade. During the proceedings, Kim revealed his dynastic ambition by appointing his son, Kim Jong Il, to three powerful party posts, thus making the younger Kim his heir apparent. The younger Kim consolidated his power and gradually assumed increasing control over the day-to-day administration of the government until his father's death in 1994. Until that time, the president held supreme power. When Kim Jong Il formally took power in 1998, he was given the new title of chairman of the National Defense Commission.

International relations. North Korea has remained one of the most isolated of all countries, with severe restrictions on travel into or out of the country, a totally controlled press, and an ideology of self-reliance. In the 1970s and '80s, the North Korean government maintained its balanced diplomatic position between the country's only two significant allies, China and the Soviet Union, while sustaining a hostile attitude toward the United States. The dissolution of the U.S.S.R. in the early 1990s left China as North Korea's sole major ally. Even China, however, could no longer be relied upon fully, as it cultivated friendly relations with South Korea from 1992.

When North Korea realized it could not count on its traditional allies to block South Korean UN membership, it retreated from its long-standing position of insisting upon a single, joint Korean seat. Both North Korea and South Korea were admitted to the UN on Sept. 17, 1991, as separate nations. Hopes for diplomatic breakthroughs with the South were dimmed, however, by anxiety over the North's nuclear weapons program. The resulting international dispute led to a negotiated settlement signed by the United States and North Korea at Geneva in 1994. Implementing the agreement proved difficult, however, and by 2003 North Korea had renounced its pledge not to build nuclear weapons.

UN membership

Relations with the South. During the late 1960s, the North had significantly escalated its activities against the South—from about 50 incidents in 1966 to more than 500 in 1967. One of its most brazen acts was an attempt in January 1968 by North Korean commandos, who penetrated to within a mile of the presidential palace, to kill South Korean President Park Chung Hee. In the early 1970s, the North ended its insurgency campaign, and talks between the North and South began at P'anmunjŏm in September 1971. High-level discussions began in early 1972—culminating in a historic joint North-South communiqué in July, in which both sides agreed to seek reunification peacefully—and continued until August 1973, when they were unilaterally suspended by the North.

Since 1980 North Korean policy toward the South has alternated, often bewilderingly, between peace overtures and provocation. In October 1980 Kim Il-sung unveiled a proposal to create a confederate republic by a loose merger of the two states, based on equal representation. Later in the decade, however, the North engineered two major terrorist incidents against the South: the first was a bombing in Rangoon, Burma (now Yangôn, Myanmar), in October 1983 that killed several members of the South Korean government; and the second, in November 1987, was the destruction by time bomb of a South Korean airliner over the border between Thailand and Burma.

North-South relations appeared to reach a milestone in 1991, as a series of prime-ministerial talks produced joint declarations in December on nonaggression and denuclearizing the Korean peninsula. In June 2000 the leaders of the two Koreas held an unprecedented meeting. However, little came of either the agreements or the later talks, especially after the North became embroiled in the controversy over its nuclear program. (Y.I.L./Ed.)

For later developments in the history of North Korea, see the BRITANNICA BOOK OF THE YEAR.

South Korea

The Republic of Korea (Korean: Taehan Min'guk), commonly called South Korea, occupies the southern portion of the Korean peninsula. The country is bordered by the Democratic People's Republic of Korea (North Korea) to the north, the East Sea (Sea of Japan) to the east, the East China Sea to the south, and the Yellow Sea to the west; to the southeast it is separated from the Japanese island of Tsushima by the Korea Strait. South Korea faces North Korea across a demilitarized zone that runs for about 150 miles (240 kilometres) roughly from the mouth of the Han River on the west coast of the Korean peninsula to a little south of the North Korean town of Kosŏng on the east coast. The area of South Korea is 38,432 square miles (99,538 square kilometres), or about 45 percent of the peninsula. The capital is Seoul (Sŏul).

PHYSICAL AND HUMAN GEOGRAPHY

The land. *Relief.* Geologically, South Korea consists in large part of Precambrian (*i.e.,* more than 540 million years old) rocks such as granite and gneiss. The country is largely mountainous, with small valleys and narrow coastal plains. The T'aebaek Mountains run in roughly a north-south direction along the eastern coastline and northward

Geologic foundation

into North Korea, forming the country's drainage divide. From them several mountain ranges branch off with a northeast-southwest orientation. The most important of these are the Sobaek Mountains, which undulate in a long S-shape across the peninsula. None of South Korea's mountains are very high: the T'aebaek Mountains reach an elevation of 5,604 feet (1,708 metres) at Mount Sŏrak, and the Sobaek Mountains reach 6,283 feet at Mount Chiri. The highest peak in South Korea, the extinct volcano Mount Halla on Cheju Island, is 6,398 feet (1,950 metres) above sea level.

South Korea has two volcanic islands—Cheju, off the peninsula's southern tip, and Ullŭng, about 85 miles east of the mainland in the East Sea—and a small-scale lava plateau in Kangwŏn province. There are fairly extensive lowlands along the lower parts of the country's main rivers. The eastern coastline is relatively straight, whereas the western and southern have extremely complicated ria (*i.e.,* creek-indented) coastlines with many islands. The shallow Yellow Sea and the complex Korean coastline produce one of the highest tidal ranges in the world—about 30 feet maximum at Inch'ŏn, the entry port for Seoul.

Drainage and soils. South Korea's three principal rivers, the Han, Kum, and Naktong, all have their sources in the T'aebaek Mountains, and they flow between the ranges before entering their lowland plains. Nearly all the country's rivers flow westward or southward into either the Yellow Sea or the East China Sea; only a few short, swift rivers drain eastward from the T'aebaek Mountains. The Naktong River, South Korea's longest, runs southward for 325 miles to the Korea Strait. Stream flow is highly variable, being greatest during the wet summer months.

Most of South Korea's soils derive from granite and gneiss. Sandy and brown-coloured soils are common, and they generally have little humus content. Podzolic soils (ash-gray forest soils), resulting from the cold of the long winter season, are found in the highlands.

Climate. The greatest influence on the climate of the Korean peninsula is its proximity to the Asian landmass. This produces the marked summer-winter temperature extremes of a continental climate while also establishing the northeast Asian monsoons (seasonal winds) that affect precipitation patterns. The annual range of temperature is greater in the north and in interior regions of the peninsula than in the south and along the coast.

Temperature

South Korea's climate is characterized by a cold, relatively dry winter and a hot, humid summer. The coldest average monthly temperatures in winter drop below freezing except along the southern coast. The average January temperature at Seoul is about 23° F (−5° C), while the corresponding average at Pusan, on the southeast coast, is 35° F (2° C). By contrast, summer temperatures are relatively uniform across the country, the average monthly temperature for August being about 77° F (25° C).

Annual precipitation ranges from about 35 to 60 inches (900 to 1,500 millimetres) on the mainland. Taegu, on the east coast, is the driest area, while the southern coast is the wettest; southern Cheju Island receives more than 70 inches annually. Up to three-fifths of the annual precipitation is received in June-August, during the summer monsoon, the annual distribution being more even in the extreme south. Occasionally, late-summer typhoons cause heavy showers and storms along the southern coast. Precipitation in winter falls mainly as snow.

Plant and animal life. The long, hot, humid summer is favourable for the development of extensive and varied vegetation. Some 4,500 plant species are known. Forests once covered about two-thirds of the total land area, but the original forest has almost disappeared. Except for evergreen broad-leaved forests in the narrow subtropical belt along the southern coast and on Cheju Island, most areas contain deciduous broad-leaved and coniferous trees. Typical evergreen broad-leaved species include the camellia and camphor tree, while deciduous forests include species of oak, maple, alder, zelkova, and birch. Species of pine are the most representative in the country; other conifers include spruces, larches, and yews.

Wild animal life is similar to that of northern and northeastern China. The most numerous larger mammals are deer. Tigers, leopards, lynx, and bears, formerly abundant, have almost disappeared, even in remote areas. Some 380 species of birds are found in the country, most of which are seasonal migrants. Many of South Korea's fish, reptile, and amphibian species are threatened by intensive cultivation and environmental pollution.

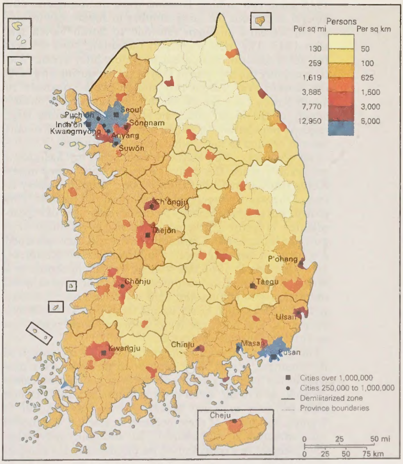
Settlement patterns. Agglomerated villages are common in river valleys and coastal lowlands in rural areas, ranging from a few houses to several hundred. Small clustered fishing villages are found along the coastline. In contrast to the lowlands, settlements in mountain areas are usually scattered. The pace of urbanization in South Korea has caused considerable depopulation of rural areas, and the traditional rural lifestyle has been slowly fading away.

Greater Seoul

In contrast to rural areas, urban populations have grown enormously. Seoul, the political, economic, and cultural centre of the country, is by far the largest city; satellite cities around Seoul—notably Anyang, Sŏngnam, Suwŏn, and Puch'ŏn—also have grown rapidly, forming an extensive conurbation (Greater Seoul) to the south of the city. The other major cities with populations of at least one million are Pusan, Taegu, Inch'ŏn, Kwangju, and Taejŏn.

The people. *Ethnic and linguistic composition.* The Korean people originally may have had links with the people of Central Asia, the Lake Baikal region, Mongolia, and the coastal areas of the Yellow Sea. Tools of Paleolithic type and other artifacts found in Sokch'ang, near Kongju, are quite similar to those of the Lake Baikal and Mongolian areas. The population of South Korea is highly homogeneous, although the number of foreigners is growing, especially in the major urban areas. In addition to American soldiers, urban Chinese, and foreign nationals in business, tens of thousands of workers have come to South Korea from China and Southeast Asia.

All Koreans speak the Korean language, which is often classified as one of the Altaic languages, has affinities to Japanese, and contains many Chinese loanwords. The Korean script, known in South Korea as Hangul (Han'gŭl) and in North Korea as Chosŏn muntcha, is composed of phonetic symbols for the 10 vowels and 14 consonants. Korean often is written as a combination of Chinese ideograms and Hangul in South Korea, although the trend there is toward using less Chinese. A number of English words and phrases also have crept into the language.



Population density of South Korea.

Religion. Freedom of religion is constitutionally guaranteed in South Korea, and there is no national religion. There also is little uniformity of religious belief. Thus, an individual may adhere to Buddhism while also following Confucian or Taoist tenets.

Buddhism, which was first introduced in the 4th century AD and was the official religion of the Koryŏ dynasty (918–1392), has the largest following. Christianity is relatively new in Korea, Roman Catholic missionaries having reached the peninsula only in the late 18th century, and their Protestant counterparts a century later. Christians now are the second largest religious group, with Protestants far outnumbering Catholics. Confucianism was the basis of national ethics during the Chosŏn (Yi) dynasty (1392–1910); though the number of its official adherents is now small, most Korean families still follow its principles, including ancestor worship. Among the so-called new religions are Wŏnbulgyo (Wŏn Buddhism), Taejongyo, and Ch'ŏndogyo. Ch'ŏndogyo ("Teaching of the Heavenly Way"), originally known as Tonghak ("Eastern Learning"), is a blend of Buddhism, Confucianism, Christianity, and even Taoism; it spread widely in the latter part of the 19th century. Shamanism—the religious belief in gods, demons, and ancestral spirits responsive to a priest, or shaman—and traditional geomancy (*p'ungsu*) persist.

Demographic trends. South Korea's population more than doubled in the period 1950–90. From 1960, however, birth and mortality rates decreased rapidly. The decrease in the birth rate was caused chiefly by a national campaign for family planning conducted since 1965. It also reflected an overall increase in living standards, considerable growth in the number of educated people, and the shift from a rural and agrarian to an urban and industrial society.

The rapid increase in the urban population and the resultant depopulation of vast rural areas are South Korea's main demographic issues. Three-fourths of the population is classed as urban, with roughly half living in the country's six largest cities. Thus, although the country's rate of population growth is low, its overall population density is high—some two and a half times that of North Korea—with huge concentrations of people in the major cities.

Population movements

Large numbers of Koreans emigrated before World War II: those from northern Korea to Manchuria, and those from southern Korea to Japan. It is estimated that in 1945 some two million Koreans lived in Manchuria and Siberia and about the same number in Japan. About half of the Koreans in Japan returned to South Korea just after 1945. The most important migration, however, was the north-to-south movement of people after World War II, especially the movement that occurred during and after the Korean War. About two million people migrated to South Korea from the North during that period, settling largely in the major cities. In addition to creating large resident populations in China and Japan, many Koreans have emigrated to the United States and Canada.

The economy. The South Korean economy has grown remarkably since the early 1960s. In that time, South Korea transformed itself from a poor, agrarian society to one of the world's most highly industrialized nations. This growth was driven primarily by the development of export-oriented industries, fostered by strong government support. Government and business leaders together fashioned a strategy of targeting specific industries for development, and beginning in 1962 this strategy was implemented in a series of economic development plans. The first targeted industries were textiles and light manufacturing, followed in the 1970s by such heavy industries as iron and steel and chemicals. Still later, the focus shifted to such enterprises as automobiles and electronics.

The government exercised strong controls on industrial development, giving most support to the large-scale projects of the emerging giant corporate conglomerates called *chaebŏl.* As a result, small and medium-size industries that were privately managed became increasingly difficult to finance, and many of these became, in essence, dependent subcontractors of the *chaebŏl.* In addition, preferential reinvestment in export industries discouraged the production of consumer goods and held down consumer spending. By 1980, however, these policies began to be reversed, especially credit policies, as the government gradually removed itself from direct involvement in industry.

Labour unions were able to win significant increases in wages during the 1980s, which improved the lot of workers and produced a corresponding growth in domestic consumption. Higher labour costs, however, contributed to a decline in international competitiveness in such labour-intensive activities as textile manufacture.

Mining and power. Mineral resources in South Korea are meagre. The most important reserves are of anthracite coal, iron ore, graphite, gold, silver, tungsten, lead, and zinc, which together constitute some two-thirds of the total value of mineral resources. Deposits of graphite and tungsten are among the largest in the world. Most mining activity centres around the extraction of coal and iron ore. All of the country's crude petroleum requirements and most of its metallic mineral needs are met by imports.

Thermal electric power accounts for more than half of the power generated. Since the first oil refinery started to produce petroleum products in 1964, power stations have changed over gradually from coal to oil. Hydroelectricity constitutes only a small proportion of overall electric-power production. Nuclear power generation, however, has become increasingly important.

Agriculture. Less than one-fourth of the republic's area is cultivated. Along with the decrease in farm population, the proportion of national income derived from agriculture has decreased to a fraction of what it was in the early 1950s. Improvements in farm productivity have been hampered because fields typically are divided into tiny plots that are cultivated largely by manual labour and animal power. In addition, the decrease and aging of the rural population has caused a serious farm-labour shortage. Rice is the most important crop, constituting about two-fifths of all farm production in value. Barley, wheat, soybeans, potatoes, millet, and a wide variety of vegetables also are important. Double-cropping of rice and barley is common in the southern provinces.

Crops

Forestry and fishing. From the 1970s successful reforestation efforts were mounted in areas previously denuded. Domestic timber production, however, supplies only a negligible fraction of demand. Logging, mainly of coniferous trees, is limited to the mountain areas of Kangwŏn and North Kyŏngsang provinces.

Fishing has long been important for supplying protein-rich foods and has emerged as a significant export source. South Korea has become one of the world's major deep-sea fishing nations. Coastal fisheries and inland aquaculture also were developed.

Industry. Textiles and other labour-intensive industries have declined from their former preeminence in the national economy. Heavy industries, including chemicals, metals, machinery, and petroleum refining, are highly developed. Industries that are even more capital- and technology-intensive grew to importance after 1980—notably shipbuilding, motor vehicles, and electronic equipment. Also from the 1980s, increasing emphasis was given to such high-technology industries as bioengineering and aerospace, and the service industry grew markedly. Much of the country's manufacturing is centred on Seoul, while heavy industry is largely based in the southeast.

Finance and trade. The government-owned Bank of Korea, headquartered in Seoul, is the country's central bank, issuing currency and overseeing all banking activity. All banks were nationalized in the early 1960s, but by the early 1990s these largely had been returned to private ownership. Foreign branch banking has been allowed in South Korea since the 1960s, and in 1992 foreigners began trading on the Korea Stock Exchange in Seoul.

South Korea borrowed heavily on international financial markets to supply capital for its industrial expansion, but the success of its exports allowed it to repay much of its debt. The major imports are machinery, mineral fuels, manufactured goods, and such crude materials as textile fibres and metal ores and scrap. Principal exports include machinery, textiles, transport equipment, and clothing and footwear. South Korea's principal trading partners are the United States, Japan, members of the European Union, and Southeast Asian countries.

Transportation. South Korea's transportation system was expanded and improved considerably after 1960, especially with the creation of a modern highway network and the establishment of nationwide air service. Road construction, however, did not keep up with the tremendous increase in the number of motor vehicles in the country, especially in urban areas. Road transport now accounts for the bulk of passenger travel and most movement of freight. The country's first multilane highway (from Seoul to Inch'ŏn) was opened in 1968, and the express highway network was expanded to link most major cities.

Express highways

The South Korean railways are largely government-owned. Until 1960 rail travel was the major means of inland transportation for both freight and passengers but since has been superseded by road transport, and, more recently, by the rise in air travel. The Seoul-Pusan line through Taejŏn and the Seoul-Inch'ŏn line are double-tracked, and many lines are electrified. Seoul and Pusan have heavily used subway systems.

Internal air transportation began in the early 1960s. Kimp'o International Airport, at Seoul, serves as the country's main port of entry, and Pusan and Cheju also have international airports. Port facilities have been expanded considerably with the tremendous growth in trade. Major ports include Pusan, Inch'ŏn, Ulsan, and Cheju.

Administration and social conditions. *Government.* The government instituted after a constitutional referendum in 1987 is known as the Sixth Republic. The government structure is patterned mainly on the presidential system of the United States and is based on separation of powers among the legislature, the executive, and the judiciary. The president, since 1987 chosen by direct popular election for a single five-year term, is the chief of state, head of the executive branch, and commander of the armed forces. The State Council, the highest executive body, is composed of the president, the prime minister, the heads of executive ministries, and ministers without portfolio. The prime minister is appointed by the president and approved by the elected National Assembly.

Legislative authority rests with the unicameral National Assembly (Kuk Hoe). The powers of the National Assembly, which was reinstated in 1980 after a period of curtailment, were strengthened in 1987. Its 299 members are chosen, as previously, by a combination of direct and indirect election to four-year terms.

South Korea is divided administratively into the nine provinces (*do* or *to*) Cheju, North Chŏlla, South Chŏlla, North Ch'ungch'ŏng, South Ch'ungch'ŏng, Kangwŏn, Kyŏnggi, North Kyŏngsang, and South Kyŏngsang; the special city (*t'ŭkpyŏlsi*) of Seoul; and the five megalopolises (*kwangyŏksi*) Pusan, Taegu, Inch'ŏn, Kwangju, and Taejŏn. Each has a popularly elected legislative council. Provincial governors and the mayors of province-level cities are popularly elected.

Political parties

South Korea had a two-party system until 1972, when the power of the pro-government party increased substantially and the activity of the opposition was restricted. During the 1980s restrictions on political parties were ended. The opposition was allowed to resume political participation, but it also tended to fragment; thus, the political system became more multiparty in character. The Democratic Justice Party (DJP; until 1981 called the Democratic Republican Party), the ruling party since its founding in 1963, was renamed the Democratic Liberal Party in 1990, following the merger of the DJP with two opposition parties. The Democratic Party and the Unification National Party have become the major opposition parties.

Justice. The judicial branch consists of the Supreme Court, three appellate courts, 12 district courts, and a family court. The Supreme Court is empowered to interpret the constitution and all other state laws and to review the legality of government regulations and activities. The chief justice is appointed by the president with the consent of the National Assembly, upon recommendation of the Judge Recommendation Council.

Armed forces and security. South Korea maintains a large, well-equipped armed-forces establishment—consisting of army, navy, and air force branches—although it is still considerably smaller than that of North Korea. The army is by far the largest component, and there is a sizable reserve force. Military service is compulsory for all males. South Korea's main military objective is to deter an attack by the North. To that end it has a Mutual Defense Treaty (1953) with the United States, and a large contingent of U.S. troops is stationed in the country.

Civilian intelligence gathering and other nonmilitary matters of national security are the responsibility of the Agency for National Security Planning (until 1980 called the Korean Central Intelligence Agency). Military intelligence is handled by the Defense Security Command. The Korean National Police combine standard police duties with responsibility for counteracting communist infiltration and controlling civil disorders.

Education. Six years of primary-school education is compulsory, and virtually all children of school age are enrolled. Most primary-school graduates go on to three years of middle school, and nearly all middle-school graduates continue on in high school or technical school. About one-third of the high-school graduates go to higher educational institutions. Graduation from a college or university grew considerably in importance in South Korea after World War II, and the number of college-level institutions increased enormously. Nearly all the most prestigious schools are located in Seoul and include the state-run Seoul National University and the private Ehwa Woman's University and Yonsei University. Admission to a school is granted through competitive entrance examinations. High-school students must endure grueling preparation work for these examinations.

Health and welfare. The availability of medical services increased enormously after the Korean War, but medical facilities and the number of personnel remain inadequate, especially in rural areas. Most people now have some sort of medical insurance coverage. The average life-expectancy rate rose dramatically from the 1950s, while the death rate more than halved.

Government welfare activities are relatively new and limited in range. The programs include care of disabled war veterans, a variety of homes (for the aged, homeless, disabled war widows, and orphans), vocational training of women, and care of juvenile delinquents. Following on the devastation of the Korean War, United Nations agencies, civilian and military agencies of the United States, and private volunteer agencies played a significant role in the steadily improving living conditions in South Korea. Also significant was the dramatic increase in household income, especially among industrial workers.

Housing. Rapid expansion of urban areas, especially the expansion of Seoul and Pusan, has resulted in considerable changes in the urban landscape. High-rise buildings, especially apartment blocks, are now common in Seoul. Because of this rapid growth, city services, such as water, transportation, and sewage systems, generally have not been able to keep up with this growth. Despite the rapid pace of urban housing construction, the shortage of housing in metropolitan areas has remained a problem, although it has been partially solved by central and local government-housing programs.

City services

Cultural life. Shamanism, Buddhism, and Confucianism constitute the background of modern Korean culture. Since World War II, and especially after the Korean War, the modern trends have rapidly progressed. Traditional thought, however, still plays an important role under the surface. Korea belongs historically to the Chinese cultural realm. After the Three Kingdoms period in particular, Korean culture was strongly influenced by the Chinese, although this influence was given a distinctive Korean stamp. The National Museum of Korea maintains artifacts of Korean culture, including many national treasures, chiefly in the central museum in Seoul; there are branch museums in eight other cities. Archaeological sites include the ancient burial mounds at Kyŏngju, capital of the Silla (Shilla) kingdom, and Kongju and Puyŏ, two of the capitals of Paekche.

Architecture. Korean architecture shows Chinese influence, but it is adapted to local conditions, utilizing wood and granite, the most abundant building materials. Beautiful examples are found in old palaces, Buddhist temples, stone tombs, and Buddhist pagodas. Western-style architecture became common from the 1970s.

Painting and ceramics. One of the earliest examples of Korean painting is found in the mural paintings in the royal tombs of Koguryŏ. The best-known mural paintings are those in the Ssangyong Tomb at Yonggang, located in North Korea. Ceramic arts became highly developed, flourishing during the Koryŏ period and diffusing to Japan, and every province continues to produce its distinctive ceramic wares.

Dance and music. Folk dances survive, and folk music, accompanied by native musical instruments, is performed occasionally at ceremonies and festive occasions. The government has made an effort to preserve the traditional arts. The National Classical Music Institute (formerly the Prince Yi Conservatory), for example, plays an important role in the preservation of folk music. The Korean National Symphony Orchestra and the Seoul Symphony Orchestra are two of the best-known organizations performing Western music.

Recreation. South Koreans are avid sports and outdoors enthusiasts. The martial art tae kwon do and the traditional wrestling style called *ssirŭm* are widely practiced national sports. There are well-supported professional baseball and association football (soccer) leagues. The 1988 Summer Olympic Games at Seoul were an enormous boost for national pride and for the enhancement of Korean cultural identity. The country's system of national parks attracts large numbers of hikers, campers, and skiers.

National parks

The press and broadcasting. Constitutionally guaranteed press freedoms, often violated before 1987, are now generally observed. There are a number of nationally distributed daily newspapers and many regional and local dailies. The Yŏnhap News Agency is the largest such organization. In addition to the publicly owned Korean Broadcasting System, a growing number of private radio and television stations have been established.

(C.Le./W.-i.Y.)

For statistical data on the land and people of South Korea, see the *Britannica World Data* section in the BRITANNICA BOOK OF THE YEAR.

HISTORY

South Korea to 1961. *The First Republic.* The First Republic, established in August 1948, adopted a presidential system and elected Syngman Rhee as the first president. He was reelected in August 1952 while the nation was at war. Even before the outbreak of the Korean War, there had been a serious conflict between Rhee and the opposition-dominated National Assembly that elected him in 1948. The dispute involved a constitutional amendment bill that the opposition introduced in an attempt to oust Rhee by replacing the presidential system with a parliamentary cabinet system. The bill was defeated.

When the opposition introduced another amendment bill in favour of a parliamentary cabinet system, Rhee in 1952 countered by pushing through a bill that provided for the popular election of the president. Rhee ran successfully for his third term of office in May 1956. Rhee's election for the fourth time, in March 1960, was preceded by a period of tension and violence. Amid massive student demonstrations, which culminated in the April 19 "student uprising" and which resulted in many casualties, Rhee resigned under pressure and fled to exile in Hawaii.

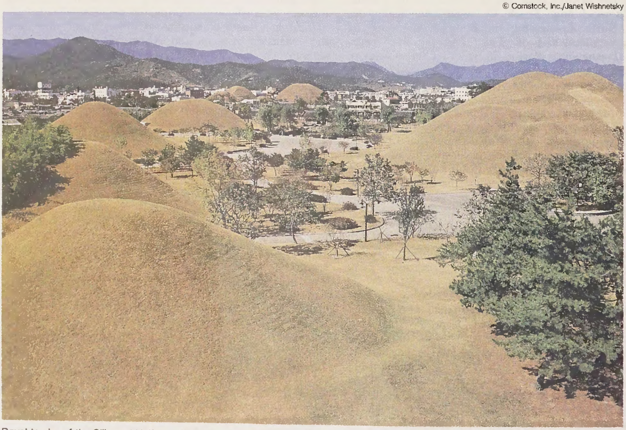
The Second Republic. The second Korean republic, which adopted a parliamentary cabinet system, lasted only nine months before it was overthrown by a military coup in May 1961. With a figurehead president elected by both houses of the legislature, power was shifted to the office of Prime Minister Chang Myŏn, who was elected by the lower house by a narrow margin of 10 votes.

The Chang regime made some strenuous efforts to initiate reforms. In a society laden with social and economic ills accumulated over a long period of time, however, it failed to cope with the unstable situation created by a violent political change. To make the situation worse, factionalism ran rampant in the political arena. With the ultimate source of authority now vested in the office of the prime minister, all factions, conservative and moderate, engaged in constant maneuvering to win over a group of independents in order to form a majority in the legislature. Before Chang had time to launch a full program of economic reform, the leadership of the ruling Democratic Party was crippled by factional strife within its ranks.

Military rule. *The 1961 coup.* With the seizure of political power by the military, on May 16, 1961, through a carefully engineered coup d'état, postliberation Korean politics entered a new phase. The military junta, led by General Park Chung Hee, took over the government machinery, dissolved the National Assembly, and imposed a strict ban on political activity. The nation was placed under martial law, and the Supreme Council for National

Martial law

© Comstock, Inc./Janet Wishnetsky



Royal tombs of the Silla and Unified Silla kingdoms (1st century BC–10th century AD) at Kyŏngju, South Korea.

Reconstruction (SCNR), headed by Park, took the reins of government and began instituting a series of reforms.

In November 1962 the SCNR made public a constitutional amendment bill that provided for a strong president and a weak, single-chamber National Assembly. The bill was approved by a national referendum one month later. Then, in February 1963, Park announced that he would not take part in the civilian government to be formed later in the year if civilian political leaders chose to uphold a nine-point "political stabilization proposal."

In March, however, as a result of bitter turbulence within the ruling junta and a chaotic situation created by the proliferation of minor political parties, Park changed his mind and proposed that military rule be extended for four years. The proposal met vigorous opposition from civilian political leaders, but some 160 military commanders, most of them generals, supported the extension. Under considerable pressure, Park announced, in April 1963, a plan for holding elections toward the end of the year. Park was named presidential candidate of the newly formed Democratic Republican Party (DRP) in late May.

The Third Republic. The election for president of the Third Republic took place on Oct. 15, 1963. Park narrowly defeated the opposition candidate, Yun Po-sŏn, former president of the Second Republic, who had remained in office as a figurehead at the request of the junta to provide constitutional continuity for the military regime. When political activity was permitted to resume, Yun led the mustering opposition groups and became the presidential candidate of the Civil Rule Party. In May 1967 Park was elected to his second term of office, and the DRP won a large majority in the National Assembly. Members of the opposition New Democratic Party (NDP), whose candidate was the twice-defeated Yun, claimed election fraud.

President Park faced the constitutional provision that limited the president to two consecutive four-year terms. Despite demonstrations by opposition politicians and students, DRP legislators passed an amendment that would allow a president to serve three consecutive four-year terms. The amendment was approved by a national referendum in October 1969. In the presidential elections held on April 27, 1971, Park defeated his opponent, Kim Dae Jung of the NDP; in elections for the National Assembly, however, the NDP made substantial gains.

The Yushin order. In December 1971, after his election to a third term, Park declared a state of emergency, and 10 months later he suspended the constitution and dissolved the legislature. A new constitution, which would permit the reelection of the president for an unlimited number of six-year terms, was promulgated in December, inaugurating the Fourth Republic.

The institutional framework of the Yushin ("Revitalization Reform") order departed radically from the Third Republic. The National Conference for Unification (NCU) was created "to pursue peaceful unification of the fatherland." The conference was to be a body of members who were directly elected by the voters for a six-year term. The president was chairman of the conference. Until 1987 the NCU was charged with the power to elect the president; and under this arrangement, Park was elected without opposition in 1972 and was reelected in 1978.

Political unrest increased following the August 1973 kidnapping from Tokyo to Seoul of Kim Dae Jung by agents of the Korean Central Intelligence Agency (KCIA). From August 1978 the opposition movement became stronger. The expulsion from the National Assembly of the new NDP leader Kim Young Sam in early October 1979 escalated what had already been growing political tensions between the government and opposition leaders during the year into a major national crisis. Antigovernment riots broke out in Pusan and were suppressed by government troops. The crisis culminated on October 26, when President Park was assassinated by his own director of the KCIA. Prime Minister Choi Kyu Hah became acting president under the Yushin constitution and was formally elected president in December by the NCU.

Assassination of Park

In the meantime, the country was placed under stern military rule by General Chun Doo Hwan. A major uprising of students and dissidents in Kwangju in May 1980, calling for the full restoration of democracy, was ruthlessly suppressed by the junta, with hundreds of civilian deaths. That month the military did away with all trappings of civilian government, extended martial law, again banned all political activity, and closed universities and colleges.

Restoration of civilian government. *The Fifth Republic.* In August 1980, Chun Doo Hwan was elected president by the NCU. A new constitution was approved in October, ushering in the Fifth Republic; the powers of the president were reduced in favour of the National Assembly, and the president was limited to one seven-year term. Martial law was lifted in January 1981, and in February Chun won election under the new constitution. As parties were again allowed to operate, a new ruling party, the Democratic Justice Party (DJP), was formed by former members of the DRP and NDP. Chun's administration, however, had to endure a series of scandals and incidents—most notably the bombing by North Koreans in Rangoon, Burma (now Yangôn, Myanmar), in October 1983 that killed several members of the South Korean government.

By 1987, popular dissatisfaction had spread. Roh Tae Woo (from 1985 chairman of the DJP) announced a constitutional-reform program to restore the democratic institutions and civil rights the military had eliminated. A revised constitution was approved by national referendum in October. Among its provisions were a reduction in the presidential term from seven to five years and the direct popular election of the president. Roh, a former army general, was elected president in December and took office in February 1988. Thus, a peaceful transfer of power was effected for the first time in South Korean history, and the Fifth Republic came to an end. (B.-h.H./Y.I.L.)

The Sixth Republic. In the much-improved political climate of the Sixth Republic, South Korea hosted the highly successful Summer Olympic Games in Seoul in 1988. Roh proceeded to bring about a merger (1990) of the DJP with the Reunification Democratic Party of Kim Young Sam and the New Democratic Republican Party of Kim Jong Pil, who for a time had been prime minister during the Fourth Republic. The resultant Democratic Liberal Party (DLP) commanded a large majority in the National Assembly. In December 1992, Kim Young Sam was elected president on the DLP ticket. He succeeded Roh in February 1993, thus becoming the first civilian president in more than three decades. Kim Dae Jung was elected president in 1997 and inaugurated in 1998. He continued the democratization process and the search for reconciliation between the two Koreas.

The Roh government pursued diplomatic initiatives with the Soviet Union and its allies, and diplomatic ties were established with Hungary, Poland, and Yugoslavia in 1989 and the Soviet Union in 1990. Relations between South Korea and China improved as well, and in 1992 the two countries established full diplomatic ties.

Relations with the North. Tensions between South Korea and the North generally remained high after the Korean War. The first significant contact between the two states occurred in early 1972, when the Park government carried out secret negotiations with North Korea. A joint statement was issued in July that announced agreement on a formula for national reunification. The ensuing dialogue between North and South, however, was short-lived.

The 1972 negotiations

In the early 1990s there were again signs of rapprochement between the two Koreas. North-South relations appeared to reach a milestone with the signing of a pact of reconciliation and nonaggression in December 1991. Earlier that year, the two states were separately admitted to the UN on September 17. North Korea's nuclear weapons capabilities emerged as a source of anxiety for the South shortly thereafter, but hope remained for further reconciliation and for a peaceful reunification of the peninsula. Kim Dae Jung's quest for reconciliation with the north led to the first-ever summit between the leaders of the two Koreas in P'yŏngyang in June 2000, for which he received the Nobel Prize for Peace in October 2000.

Economic and social developments. South Korea in the 1950s had an underdeveloped, agrarian economy that was dependent on foreign aid. The military leadership that emerged in the early 1960s and led the country for a quarter century was autocratic and, at times, repressive, but its commitment to economic development resulted in what became known as the "miracle on the Han River." South Korea became an industrial powerhouse with a skilled labour force. The country was hit hard by the Asian economic crisis of the late 1990s, but it has been restructuring its economy and recovering rapidly.

South Korean society underwent an equally rapid transformation. The population more than doubled in size between 1953 and 2000. Simultaneously, modern education developed rapidly, again with considerable government involvement but also because of the resurgence of the Korean people's zeal for education after decades of repression under Japanese rule. The growth of educational institutions and of commercial and industrial enterprises in and around South Korea's major cities attracted an increasing number of rural people to urban areas. An ambitious program also was undertaken to expand and modernize the country's transportation infrastructure.

The most conspicuous social change in South Korea, however, was the emergence of a middle class. Land-reform policies carried out in the early 1950s, together with the spread of modern education and the expansion of the economy, eliminated the once-privileged *yangban* (landlord) class, and a new elite emerged from the ranks of the commoners. Another social change was the decline of the extended family as more urban dwellers lived in apartments as nuclear families and had fewer children. In addition, women campaigned for legal equality and won enhanced property ownership rights and the right to become household heads. (Y.I.L./Ed.)

For later developments in the history of South Korea, see the BRITANNICA BOOK OF THE YEAR.

BIBLIOGRAPHY

History. Broad overviews include KI-BAIK LEE (KI-BAEK YI), *A New History of Korea* (1984; originally published in Korean, 1966), which begins with prehistory and ends with the April 1960 revolution; ANDREW C. NAHM, *Korea: Tradition & Transformation*, 2nd ed. (1996), and *Introduction to Korean History and Culture* (1993, revised reprint, 1994), both with emphasis on the modern period covering North and South Korea; CARTER J. ECKERT *et al.*, *Korea, Old and New* (1990); and ROBERT T. OLIVER, *A History of the Korean People in Modern Times: 1800 to the Present* (1993).

Korea's early history is detailed in MARTINA DEUCHLER, *The Confucian Transformation of Korea* (1992). Various periods and events in 19th-century Korea are dealt with in KEY-HIUK KIM, *The Last Phase of the East Asian World Order: Korea, Japan, and the Chinese Empire, 1860–1882* (1980); JAMES B. PALAIS, *Politics and Policy in Traditional Korea* (1975, reissued 1991), a thorough analysis of Yi dynasty politics and economy covering the period 1864–76; MARTINA DEUCHLER, *Confucian Gentlemen and Barbarian Envoys: The Opening of Korea, 1875–1885* (1977); GEORGE ALEXANDER LENSEN, *Balance of Intrigue: International Rivalry in Korea & Manchuria, 1884–1899*, 2 vol. (1982); and VIPAN CHANDRA, *Imperialism, Resistance, and Reform in Late Nineteenth-Century Korea* (1988).

The country's history in the 20th century is chronicled in CHONG-SIK LEE, *The Politics of Korean Nationalism* (1963), a comprehensive study of the nationalist movement during Japanese colonial rule; ANDREW J. GRAJDANZEV (ANDREW J. GRAD), *Modern Korea* (1944, reprinted 1978), a critique of Japanese imperial rule in Korea; CARTER J. ECKERT, *Offspring of Empire: The Koch'ang Kims and the Colonial Origins of Korean Capitalism, 1876–1945* (1991), a well-written study arguing that Korean capitalism first arose during the colonial period; MICHAEL EDSON ROBINSON, *Cultural Nationalism in Colonial Korea, 1920–1925* (1988); SOON SUNG CHO, *Korea in World Politics, 1940–1950: An Evaluation of American Responsibility* (1967); and BRUCE CUMINGS, *The Origins of the Korean War*, 2 vol. (1981–90), a most influential work on the background of the Korean War from a revisionist viewpoint. DAVID REES, *Korea: The Limited War* (1964, reissued 1978), is one of the best books dealing with the war. (Y.I.L.)

The modern states of Korea. DONALD STONE MACDONALD, *The Koreans: Contemporary Politics and Society*, 3rd ed. (1996), covers geography, history, culture, and economics. LEE-JAY CHO and YOON HYUNG KIM (eds.), *Economic Systems in South and North Korea: The Agenda for Economic Integration* (1995), provides a comparative overview of the two economies.

Traditional attitudes, customs, and values in Korea are outlined in PAUL S. CRANE, *Korean Patterns*, 4th ed., rev. (1978). HAGAN KOO (ed.), *State and Society in Contemporary Korea* (1993), discusses the social movements of North and South Korea. Women's roles are studied by YUNG-CHUNG KIM (ed. and trans.), *Women of Korea: A History from Ancient Times to 1945*, trans. from Korean (1976, reissued 1982); and SANDRA MATTIELLI (ed.), *Virtues in Conflict: Tradition and the Korean Woman Today* (1977). JON CARTER COVELL, *Korea's Cultural Roots*, 7th ed. (1985), is an introduction; while TAE HUNG HA, *Guide to Korean Culture* (1968, reissued 1985), surveys the varied phases of Korean culture. Comprehensive treatments of all Korean arts include EVELYN MCCUNE, *The Arts of Korea* (1962); CHEWŎN KIM and LENA KIM LEE (I-NA KIM), *Arts of Korea* (1974), and *The Arts of Korea*, 6 vol. (1979).

Works on Korean economic history include NORMAN JACOBS, *The Korean Road to Modernization and Development* (1985), which deals with economic history beginning with imperial Korea. The political climate of the peninsula is surveyed in SUNG CHUL YANG, *The North and South Korean Political Systems*, rev. ed. (1999); YOUNG WHAN KIHL, *Politics and Policies in Divided Korea* (1984), an informative comparative study of North and South Korean political systems; RALPH N. CLOUGH, *Embattled Korea: The Rivalry for International Support* (1987); and EUI-GAK HWANG, *The Korean Economies: A Comparison of North and South* (1993).

North Korea: ANDREA MATLES SAVADA (ed.), *North Korea, a Country Study*, 4th ed. (1994); and TAI SUNG AN, *North Korea: A Political Handbook* (1983), are objective studies of all aspects of the country. ROBERT A. SCALAPINO (ed.), *North Korea Today* (1963), is a collection of essays on political, social, and economic development in the early postwar period.

Politics and history are addressed in ROBERT A. SCALAPINO and CHONG-SIK LEE, *Communism in Korea*, 2 vol. (1972); CHIN O. CHUNG (CHIN-WI CHŎNG), *P'yongyang Between Peking and Moscow: North Korea's Involvement in the Sino-Soviet Dispute, 1958–1975* (1978); ROBERT SCALAPINO and JUN-YOP KIM (CHUN-YŎP KIM) (eds.), *North Korea Today: Strategic and Domestic Issues* (1983); and CHONG-SIK LEE and SE-HEE YOO (eds.), *North Korea in Transition* (1991). THOMAS H. HENRIKSEN and JONGRYN MO (eds.), *North Korea after Kim Il Sung* (1997); and DAE-SOOK SUH and CHAE-JIN LEE (eds.), *North Korea after Kim Il Sung* (1998), discuss political and economic development and external relations, especially U.S.–North Korea relations and North Korea's nuclear weapons program. JAE H. KU and TAE HWAN OK (eds.), *Change and Challenge on the Korean Peninsula: Developments, Trends, and Issues* (1996), examines the North–South Korea and U.S.–South Korea relationships in detail. CHONG BONG-UK (ed.), *North Korea: Uneasy, Shaky Kim Jong-Il Regime* (1997), introduces the political and socioeconomic circumstances associated with the regime of Kim Jong Il and discusses change in North Korea. HELEN-LOUISE HUNTER, *Kim Il-Song's North Korea* (1999), describes Kim Il Sung's life and social conditions in North Korea. JAE KYU PARK (ed.), *North Korea in Transition and Policy Choices* (1999), provides a description of political, economic, and social structure in transition.

South Korea: ANDREA MATLES SAVADA and WILLIAM SHAW (eds.), *South Korea, a Country Study*, 4th ed. (1992), is a good general source. LEE CHAN (CH'AN YI) *et al.*, *Korea: Geographical Perspectives* (1988), is also useful. *Korea Annual* compiles chronologies, history, statistics, and yearly highlights, with an emphasis on South Korea. KOREAN OVERSEAS INFORMATION SERVICE, *A Handbook of Korea*, 10th ed. (1998), also focuses on South Korea. Two analyses of Korean religious life are ROGER L. JANELLI and DAWNHEE YIM JANELLI, *Ancestor Worship and Korean Society* (1982); and DONALD N. CLARK, *Christianity in Modern Korea* (1986).

Studies of South Korea's economic development include DENNIS L. MCNAMARA, *The Colonial Origins of Korean Enterprise, 1910–1945* (1990); EDWARD S. MASON *et al.*, *The Economic and Social Modernization of the Republic of Korea* (1980); ALICE H. AMSDEN, *Asia's Next Giant: South Korea and Late Industrialization* (1989); RICHARD M. STEERS, YOO KEUN SHIN (YU-GŬN SIN), and GERARDO R. UNGSON, *The Chaebol* (1989); BYUNG-NAK SONG (PYŎNG-NAK SONG), *The Rise of the Korean Economy*, 3rd ed. (2001); IL SAKONG, *Korea in the World Economy* (1993), by a former finance minister; CHO SOON (SOON CHO), *The Dynamics of Korean Economic Development* (1994); and a work edited by SUNG YEUNG KWACK (SŬNG-YŎNG KWACK), *The Korean Economy at a Crossroad* (1994). DONG-SE CHA *et al.* (eds.), *The Korean Economy 1945–1995: Performance and Vision for the 21st Century* (1997), focuses on the socioeconomic development and future prospects of South Korea. JOHN KIE-CHANG OH, *Korean Politics: the Quest for Democratization and Economic Development* (1999), explores the historical processes of political and economic development.

Studies of South Korea's history include DONALD N. CLARK (ed.), *The Kwanju Uprising: Shadows Over the Regime in South Korea* (1988); HAROLD C. HINTON, *Korea Under New Leadership: The Fifth Republic* (1983); and FRANK GIBNEY, *Korea's Quiet Revolution* (1992). (W.-i.Y./Ed.)

Korean Literature

Although Korea has had its own language for several thousand years, it has had a writing system only since the mid-15th century, when the Korean alphabet, Han'gŭl, was invented. As a result, early literary activity was in Chinese characters. Korean scholars were writing poetry in the traditional manner of classical Chinese at least by the 4th century AD. A national academy was established shortly after the founding of the Unified Silla dynasty (668–935); and, from the institution of civil-service examinations in the mid-10th century until their abolition in 1894, every educated Korean had read the Confucian Classics and Chinese histories and literature. The Korean upper classes were therefore bilingual in a special sense: they spoke Korean but wrote in Chinese.

By the 7th century a system, called *idu*, had been devised that allowed Koreans to make rough transliterations of Chinese texts. Eventually, certain Chinese characters were used for their phonetic value to represent Korean particles of speech and inflectional endings. A more extended system of transcription, called *hyangch'al*, followed shortly thereafter, in which entire sentences in Korean could be written in Chinese. In another system, *kugyŏl*, abridged versions of Chinese characters were used to denote grammatical elements and were inserted into texts during transcription. Extant literary works indicate, however, that before the 20th century much of Korean literature was written in Chinese rather than in Korean, even after the invention of Han'gŭl.

In general, then, literature written in Korea falls into three categories: works written in the early transcription systems, those written in Han'gŭl (Hankul), and those written in classical Chinese.

This article is divided into the following sections:

TRADITIONAL FORMS AND GENRES

Poetry. There are four major traditional poetic forms: *hyangga* ("native songs"); *pyŏlgok* ("special songs"), or *changga* ("long poems"); *sijo* ("current melodies"); and *kasa* ("verses"). Other poetic forms that flourished briefly include the *kyŏnggi*-style, in the 14th and 15th centuries, and the *akchang* ("words for songs") in the 15th century. The most representative *akchang* is *Yongbi ŏch'ŏn ka* (1445–47; *Songs of Flying Dragons*), a cycle compiled in praise of the founding of the Yi dynasty. Korean poetry originally was meant to be sung, and its forms and styles reflect its melodic origins. The basis of its prosody is a line of alternating groups of three or four syllables, which is probably the most natural rhythm to the language.

The oldest poetic form is the *hyangga*, poems transcribed in the *hyangch'al* system, dating from the middle period of the Unified Silla dynasty to the early period of the Koryŏ dynasty (935–1392). The poems were written in four, eight, or 10 lines; the 10-line form—comprising two four-line stanzas and a concluding two-line stanza—was the most popular. The poets were either Buddhist monks or members of the Hwarangdo, a school in which chivalrous youth were trained in civil and military virtues in preparation for state service. Seventeen of the 25 extant *hyangga* are Buddhist in inspiration and content.

The *pyŏlgok*, or *changga*, flourished during the middle and late Koryŏ dynasty. It is characterized by a refrain either in the middle or at the end of each stanza. The refrain establishes a mood or tone that carries the melody and spirit of the poem or links a poem composed of discrete parts with differing contents. The theme of most of these anonymous poems is love, the joys and torments of which are expressed in frank and powerful language. The poems were sung to musical accompaniments chiefly by women entertainers, known as *kisaeng*.

Sijo and *kasa*

The *sijo* is the longest enduring and most popular form of Korean poetry. Although some poems are attributed to writers of the late Koryŏ dynasty, the *sijo* is primarily a poetic form of the Yi dynasty (1392–1910). *Sijo* were still being written in the second half of the 20th century. They are three-line poems in which each line has 14 to 16 syllables, and the total number of syllables seldom exceeds 45. Each line consists of groups of four syllables. *Sijo* may deal with Confucian ethical values, but there are also many poems about nature and love. The principal writers of *sijo* in the first half of the Yi dynasty were members of the Confucian upper class (*yangban*) and *kisaeng*. In the latter part of the Yi dynasty, a longer form, called *sasŏl sijo* ("narrative *sijo*"), evolved. The writers of this form were mainly common people; hence, the subject matter included more down-to-earth topics, such as trade and corruption, as well as the traditional topic of love. In addition, *sasŏl sijo* frequently employed slang, vulgar language, and onomatopoeia.

The *kasa* developed at about the same time as the *sijo*. In its formative stage, *kasa* borrowed the form of the Chinese *tz'u* (lyric poetry) or *fu* (rhymed prose). The *kasa* tends to be much longer than other forms of Korean poetry and is usually written in balanced couplets. Either line of a couplet is divided into two groups, the first having three or four syllables and the second having four syllables. The history of the *kasa* is divided into two periods, the division being marked by the Japanese invasion of 1592–97. During the earlier period the poem was generally about 100 lines long and dealt with such subjects as female beauty, war, and seclusion. The writers were usually *yangban*. During the later period the poem tended to be longer and to concern itself with moral instruction, travel accounts, banishment, and the writer's personal misfortunes. The later writers were usually commoners.

Akchang poems

Immediately after the founding of the Yi dynasty at the end of the 14th century and the establishment of the new capital in Seoul, a small group of poetic songs called *akchang* was written to celebrate the beginning of the new dynasty. In its earliest examples the form of *akchang* was comparatively free, borrowing its style from early Chinese classical poetry. Whereas the early *akchang* are generally short, the later *Yongbi ŏch'ŏn ka* consists of 125 cantos.

Prose. Korean prose literature can be divided into narratives, fiction, and literary miscellany. Narratives include myths, legends, and folktales found in the written records. The principal sources of these narratives are the two great historical records compiled during the Koryŏ dynasty: *Samguk sagi* (1146; "Historical Record of the Three Kingdoms") and *Samguk yusa* (1285; "Memorabilia of the Three Kingdoms"). The most important myths are those concerning the Sun and the Moon, the founding of Korea by Tangun, and the lives of the ancient kings. The legends touch on place and personal names and natural phenomena. The folktales include stories about animals; ogres, goblins, and other supernatural beings; kindness rewarded and evil punished; and cleverness and stupidity. Because the compiler of the *Samguk yusa* was a Zen master, his collection includes the lives of Buddhist saints; the origin of monasteries, stupas, and bells; accounts of miracles performed by Buddhas and bodhisattvas; and other tales rich in shamanist and Buddhist elements. The compilations made in the Koryŏ period preserved the stories of prehistoric times, of the Three Kingdoms, and of the Silla dynasty and have remained the basic sources for such material. Later compilations made during the Yi dynasty served as a major source of materials for later Yi dynasty fiction.

Korean fiction can be classified in various ways. First, there is fiction written in Chinese and that written in Korean. Second, there are the short works of one volume, "medium" works of about 10 volumes, and long works of more than 10 volumes. Third, there are works of *yangban* writers and those of common writers. In respect to the last classification, however, there is also a group of fictional works in which the viewpoints of the *yangban* and the commoner are combined. Most of this fiction was based on the narratives mentioned above, the author adding incidents and characters to the original story. It is not possible to assign definite dates or authors to most of these works. The stories are generally didactic, emphasizing correct moral conduct, and almost always have happy endings. Another general characteristic is that the narratives written by *yangban* authors are set in China, whereas those written by commoners are set in Korea.

The literary miscellany consists of random jottings by the *yangban* on four broad topics: history, biography, autobiography, and poetic criticism. Like fiction, these jottings were considered to be outside of the realm of officially sanctioned Chinese prose (*e.g.,* memorials, eulogies, and records), but they provided the *yangban* with an outlet for personal expression. Thus, their portrayal of the customs, manners, and spirit of the times in which they were composed make these writings an essential part of Korean prose.

Oral literature. Oral literature includes all texts that were orally transmitted from generation to generation until the invention of Han'gŭl—ballads, legends, mask plays, puppet-show texts, and *p'ansori* ("story singing") texts.

In spite of the highly developed literary activity from early in Korean history, song lyrics were not recorded until the invention of Han'gŭl. These orally transmitted texts are categorized as ballads and are classified according to singer (male or female), subject matter (prayer, labour, leisure), and regional singing style (capital area, western, and southern). The songs of many living performers, some of whom have been designated as "intangible national treasures" by the South Korean government, are still being recorded.

Legends include all those folk stories handed down orally and not recorded in any of the written records. These legends were for long the principal form of literary entertainment enjoyed by the common people. They deal with personified animals, elaborate tricks, the participation of the gods in human affairs, and the origin of the universe.

Mask plays

The mask plays are found in Hahoe, Chinju, T'ongyŏng, Kimhae, and Tongnae in North and South Kyŏngsang provinces; Yangju in Kyŏnggi Province; Pongsan in Hwanghae Province; and Pukch'ŏng in south Hamgyŏng Province. The most representative plays are the *sandae kŭk* genre of Yangju, the *pyŏlsin kut* of Hahoe, and the *okwangdae nori* (five-actor play) of Chinju. Although the origin of these plays is uncertain, they are generally presumed to have developed from primitive communal ceremonies. Gradually, the ceremonial aspect of the plays disappeared, and their dramatic and comic possibilities were exploited. The dialogue was somewhat flexible, the actors being free to improvise and satirize as the occasion demanded. The plays were not performed on a stage, and there were no precise limits as to the space or time in which the performances took place. The audience also traditionally responded vocally to the play as well as passively watching it. The organization of the mask plays—through repetition and variety—achieves a remarkable effect of dramatic unity.

Only two puppet-show texts are extant, *Kkoktukaksi nori* (also called *Pak Ch'ŏmjikuk;* "Old Pak's Play") and *Mansŏk chung nori.* Both titles are derived from names of characters in the plays. No theory has been formulated as to the origin and development of these plays. The plots of the puppet plays, like those of the mask plays, are full of satiric social criticism. The characters—Pak Ch'ŏmji, governor of P'yŏngam, Kkoktukaksi, Buddhist monk, and Hong Tongji—dance and sing, enacting familiar tales that expose the malfeasance of the ruling classes.

The final type of folk literature is found in the texts of *p'ansori* of the Yi dynasty. These texts were first recorded in the 19th century as verse, but the written forms were later expanded into *p'ansori* fiction, widely read among the common people. This transformation from poetry to narrative fiction was easily accomplished, since *p'ansori* were always narrative. Originally the entire *p'ansori* performance repertoire consisted of 12 *madang* ("titles"). Although all 12 remain as narrative fiction, only five of them are sung today. The texts evolved gradually from the legends, which provided their sources and were altered and expanded as they were passed from one performer to another.

HISTORY

The earliest literature: before 57 BC. From the earliest times, poetry and music have played an important part in the daily life of the Korean people. This love for song and dance impressed the ancient Chinese, whose observations are found in their early records. Ancient Korean songs, closely allied to the religious life of the people, were performed at such rites as the worship of heaven in the north and the sowing and harvest festivals in the south. These songs were transmitted orally and were thought to have magical properties.

Early songs

Three songs are handed down in Chinese translation: "Kuji ka" (or "Yŏng singun ka"; "Song for Welcoming the Gods," in the *Samguk yusa*), "Hwangjo ka" (17 BC; "Song of Orioles," in the *Samguk sagi*), and "Kong mudoha ka" (or "Konghuin"; "A Medley for the Harp," in the *Haedong yŏksa*). The "Kuji ka" is related to the myth of the founding of the Karak state, but it appears to have been a prayer sung at shamanist rituals. Some have interpreted it as being a song of seduction sung by women. The "Hwangjo ka," attributed to King Yuri, seems to be a fragment of a love song. The hero of "Kong mudoha ka" is thought to have been a shaman who drowned himself while in a trance. Perhaps the poem indicates the loss of the shaman's efficacy and authority when ancient Korea was transformed into a structured state. The story also includes other characters such as the sailor, his wife, and her friend. Another song, the "Tosol ka" (AD 28), is mentioned in the *Samguk sagi* as the beginning of secular poetry, but the poem itself has not survived.

Literature of the Three Kingdoms: 57 BC–AD 668. In contrast to the literature of the earliest ages, which is characterized by collective artistic activity, that of later ages shows the effects of political, economic, and cultural changes as the peninsula increased in wealth and widened its contacts with other areas. The introduction of Buddhism and Chinese characters to the Three Kingdoms enriched their literature and changed their worldview greatly. In consequence, their artistic activity advanced far beyond collective singing and dancing to the direct expression of individual feelings. The heroes of this literature were human beings with individual personalities in contrast to the more idealized tribal heroes of earlier times.

The three kingdoms of this period were Koguryŏ, in the north; Paekche, in the southwest; and Silla, in the southeast. The writers of Koguryŏ, the geographical location of which provided close contact with the Chinese mainland, seem to have retained something of the original pioneer spirit from the times when Koreans came from the northern regions and settled on the peninsula; their poems tended to be heroic tales in epic form. The foundation myth of Koguryŏ concerns the migration of King Tongmyŏng and his people into the region. The stories of Ondal, King Mich'ŏn, Prince Hodong, the heir apparent Yuri, and others that had their origin in Koguryŏ are still used today as the bases for dramas and motion pictures.

In contrast to that of Koguryŏ, the literature of Paekche and Silla tended to be lyrical, perhaps because of the milder climate and easier life in the south. Although little literature from Paekche has survived, the legends and songs contained in the *Samguk sagi* give a hint of its original extent and richness. For example, "Chŏngŭpsa" ("Song of Chŏngŭp")—in which the wife of an itinerant merchant asks the Moon to protect her husband—was passed down from Paekche through the Koryŏ and Yi dynasties and is still appreciated in the 20th century.

Silla led the other two kingdoms both politically (as proved by its subsequent unification of Korea) and artistically, in spite of the fact that it was farthest removed from contact with Chinese culture. The geographical and cultural distance from China, however, seems to have been an advantage, since the culture of Silla was able to create a true synthesis of native and foreign elements.

Absorption of Chinese culture

Literature of Unified Silla: 668–935. After the mid-7th century Silla absorbed Koguryŏ and Paekche and created a stable political system covering most of the Korean peninsula. During the Unified Silla dynasty many students were sent at government expense to study in T'ang China. The consequent absorption of Chinese culture and the flourishing of Korean Buddhism both contributed to the remarkable artistic flowering of Silla. In particular, the spiritual life of the Silla nobility—the monks and the chivalrous Hwarangdo—was dominated by Buddhism, and Buddhism thus became the driving force behind virtually all artistic activity.

The *hyangga* was the crown of Silla's literary achievement. Although the term *hyangga* is used generally to distinguish Korean songs from Chinese poetry, it more specifically denotes the 25 extant poems transcribed in the newly devised *hyangch'al* system in the Unified Silla and early Koryŏ periods. The texts of 14 *hyangga* are preserved in the *Samguk yusa* and those of 11 devotional poems by the Buddhist monk Kyunyŏ in *Kyunyŏ chŏn* (1075; "Life of Kyunyŏ"). A large collection, *Samdaemok,* compiled by the monks Taegu and Wi Hong in 888, has not survived. The poems that remain reveal a delicate and elegant style. Two examples written in the 8th century include "Ch'an Kip'arang ka" ("Ode to the Knight Kip'a"), which praises a member of the Hwarangdo, and "Che mangmae ka" ("Song of Offerings to a Deceased Sister"), a funeral hymn.

At the same time, a great body of prose narratives was also being written in classical Chinese. These include hundreds of volumes of commentaries on Buddhist scriptures by such monks as Wŏnhyo, Ŭisang, Wŏnch'ŭk, Taehyŏn, and Kyŏnghŭng; stories of miracles performed by eminent monks, tales of the efficacy of Buddhist statues, and the origins of Buddhist monasteries; stories of valour by members of the Hwarangdo; and stories inspired by the Chinese narrative form *ch'uan-ch'i* ("tales of marvels"). The last three types of narratives in particular became the basis of classical fiction in later dynasties.

Literature of Koryŏ: 935–1392. The last master of the *hyangga* was the monk Kyunyŏ, who wrote voluminous commentaries on, and was a great popularizer of, Buddhism. He composed his poems in Korean, transmitted them orally, and encouraged his followers to chant and memorize them. The poems in his *Kyunyŏ chŏn,* based on the 10 vows of the bodhisattva Samantabhadra, were transcribed from this oral transmission. The new poetic form that flourished during the period was the *pyŏlgok,* which was of folk origin. The *pyŏlgok* was intended for large-scale performances on festive occasions, especially the Harvest Festival and the Lantern Festival. Many *pyŏlgok* were written and performed by women, and such poems as "Tongdong" ("Ode on the Seasons") and "Isanggok" ("Winter Night") are among the most moving love lyrics in the Korean language.

The Koryŏ dynasty was a time of social instability. Internal and external crises abounded, the result of a factious and oppressive nobility and army, constant border harassment by the Khitan and Juchen peoples, and the invasions of the Mongols. Under such conditions established scholarly writers tended to be introspective or hedonistic. Consequently, the new intellectuals who arose toward the end of the dynasty began to adopt Confucian and Taoist dualistic thought as their philosophy. They were dissatisfied with *pyŏlgok* and sought a different form of poetic expression. This was the genesis of the *sijo,* which became a popular poetic form in the Yi dynasty.

Prose narratives underwent much development during the Koryŏ period. These included myths, legends, folklore, Buddhist stories and lives of saints, and literary miscellany. One notable class of tales is that in which the hero is represented by a personified inanimate object, such as wine, paper, a cane, ice, or a coin, or by an animate object, such as bamboo or a turtle. Representative of this form is *Kongbang chŏn* ("Tale of the Square-Holed Coin"), by Im Ch'un. Another major style is heroic narrative poetry, of which the masterpiece is the "Tongmyŏng wang p'yŏn" (1193; "Lay of King Tongmyŏng"), by Yi Kyubo, written in an old, pentasyllabic style. A work in a similar vein is the *Chewang ungi* (1287; "Rhymed History of Emperors and Kings"), by Yi Sŭnghyu, written in lines of five and seven syllables. A notable example of hagiography is the *Haedong kosŭng chŏn* (1215; "Lives of Eminent Korean Monks"), by Kakhun. The first collection of essays on poetry and other current subjects written in Korea is the *P'ahan chip* (1260; "Jottings to Break Up Idleness"), by Yi Inno (or Yi Illo). In addition to poetic criticism, the random jottings of Yi Inno contain autobiographical information in diary form; biographical notes on his friends and associates, including their life-styles and literary tastes; and remarks on contemporary manners and mores. The *P'ahan chip* inaugurated a long tradition of similar works written in the late Koryŏ and Yi dynasties.

Literature of the early Chosŏn period: 1392–1598. The literature of the Yi dynasty falls naturally into two periods, with the end of the Japanese invasion (1597) serving as a dividing line. The early period is notable for its poetry; the later, for its prose. Inheriting the tradition of Silla and Koryŏ, the writers of the early Yi dynasty raised Korean literature to new heights.

Invention of Korean alphabet

The early Yi dynasty also marks the initiation of a new era in Korean literary history with the invention of Han'gŭl in 1443–44, during the reign of King Sejong. This important event finally enabled Korean writers to record works in their native language.

The extraordinary king Sejong was not only the motivating force behind the invention of Han'gŭl but also had his scholars compile *Yongbi ŏch'ŏn ka* to praise the founding of the Yi dynasty, especially the valour and virtue of his father and grandfather. He himself compiled *Wŏrin ch'ŏngang chigok* (1447; "Songs of the Moon's Reflection on a Thousand Rivers") in praise of the life of the Buddha. Both works helped test and demonstrate the practicality of Han'gŭl as a means of literary expression and were the prototype of the new *akchang* form. Scholar-officials used the form to justify the founding of the new dynasty and to praise the virtues of its founder and the beauty of the new capital. As a literature of the privileged class, the popularity of the *akchang* was always limited, and it was soon eclipsed by the most important forms of the Yi dynasty—*sijo* and *kasa.*

These forms owe their popularity to two factors. First, their style of expression was rich and natural and was widely appreciated by readers. Second, they were popular with writers because together the forms provided ideal outlets for the two sides of the Confucian temper: the brief and simple *sijo* were perfect vehicles for intense lyrical expression, whereas the longer *kasa* gave writers an opportunity to expound at greater length on the more practical aspects of Confucian thought.

The expressive content of the *sijo* ranges from the idealistic union of man and nature (often coupled with the poet's pride in his poverty) to the longing for sovereigns by subjects in exile (allegorical pieces in which an analogy is drawn between fidelity and romantic love) to the deeper exploration of human problems. Writers of *sijo* include Maeng Sasŏng, Yi Hyŏnbo, Yi Hwang, and Yi I. Representative poets of *kasa* include Chŏng Ch'ŏl and Hŏ Nansŏrhŏn.

Even after the invention of Han'gŭl, prose continued to be written in Chinese. The five stories contained in the *Kŭmo sinwha* ("New Stories from Golden Turtle Mountain") by Kim Sisŭp, for example, are in the tradition of the *ch'uan-ch'i*. Subject material includes love affairs between mortals and ghosts and dream journeys to the underworld or to the Dragon Palace. Two collections of literary miscellany, the *P'aegwan chapki* ("The Storyteller's Miscellany") by Ŏ Sukkwŏn and the *Yongjae ch'onghwa* ("Miscellany of Yongjae") by Sŏng Hyŏn, were written in Chinese and influenced the growth and development of vernacular prose in the later Yi dynasty.

Shift from poetry to prose

Literature of the later Chosŏn period: 1598–1894. The shift in emphasis from poetry to prose after the Japanese invasion represents a significant step in the evolution toward modern literature. It also reflects a basic change in the philosophical outlook of Korean society. The Yi dynasty had suffered from the rigid formalism of Confucian officials, whose doctrine was based on the principles of the 12th-century Chinese philosopher Chu Hsi. This Neo-Confucian philosophy was gradually replaced by the Sirhak, or Silhak ("Practical Learning"), school, which was based on reason and the scientific spirit of criticism. The introduction of Roman Catholicism from the West and of new scientific ideas from China also stimulated the reform measures advocated by the champions of the new school.

Practical Learning gave impetus to literary activity and awakened the self-consciousness of the common people. Poetry, which had been the monopoly of the lettered class, came to be written by the common people. Women also were admitted into the literary world as the principal audience for traditional fiction. The later active compilation of *sijo* and prose narratives reveals the awakening interest in rediscovering and reappraising the past.

Prose. The traditional vernacular fiction—commonly called *sosŏl* ("small talk")—that emerged during this period consisted of stories, romances, and fables. The 15th-century *Kŭmo sinwha,* written in Chinese, was an important precursor, but the first work of the genre was *Hong Kiltong chŏn* ("Tale of Hong Kiltong"), written in the early 17th century by the scholar Hŏ Kyun. Kim Manjung, building on this style, wrote two major works: *Kuun mong* (1687–88; "Dream of Nine Clouds"), the story of a Buddhist monk's search for Enlightenment, and *Sassi namjŏng ki* (*c.* 1689–92; "Story of Lady Sa's Journey to the South"), a satire against the institution of concubinage. The most popular stories of the 18th century were all anonymous: *Ch'unhyang chŏn* ("Story of Spring Fragrance"), *Shim Ch'ŏng chŏn* ("Story of Shim Ch'ŏng"), *Changhwa hongnyŏn chŏn* ("Tale of Rose Flower and Pink Lotus"), and *Hŭngbu chŏn* ("Story of Hŭngbu"). These stories were written in a simple and natural style, their characters being modeled on common people, and they have become deeply rooted in Korean consciousness.

Court literature

Stories set at court and written by women also flourished during this period. Memorable works of court literature include the *Hanjung nok* (1795–1805; "Record of Sorrowful Days"), the tragic story of a succession dispute written by Lady Hong, princess of Hyegyŏng Palace; *Kyech'uk ilgi* ("The Diary of Kyech'uk"), the anonymous record of Queen Inmok's confinement after the assassination of her son; and *Inhyŏn wanghu chŏn* ("Tale of Queen Inhyŏn"), an anonymous account of the rivalry between the Queen and the King's concubine. All three of these works described events that had actually taken place. Other prose works written by women in Han'gŭl include diaries, travel records, letters, and portraits. These works, written in prose that verged on lyricism, could easily be chanted and memorized by a growing female readership.

Poetry. During the later Yi dynasty there was also a great flowering of poetry by scholar-officials and commoners. The most gifted poet of the period was Yun Sŏndo. His 77 *sijo* poems, including *Ŏbu sasi sa* (1651; *The Angler's Calendar*), a cycle of 40 poems on the theme of the fisherman as sage, show his mastery of topics and techniques of the *sijo*. Gradually, the *sijo* was superseded in popularity by the *sasŏl sijo*. The growth of this new form, together with the rise of fiction, drama, genre painting, and *p'ansori,* reflects the rise of the middle class and changes in the approach to life.

Pak Inno, the master of *kasa* in the 17th century, wrote in a style that combined erudition and lyricism. He produced seven pieces between 1598 and 1636; the theme of his first two *kasa* was the Japanese invasion, during which he served in the navy. The desire to reevaluate the past and to re-create the world of literature led to changes in the *kasa,* as exemplified in the anonymous *kasa* by women and commoners. Women writers of *kasa* were mainly from the southern regions of Korea. They expressed their joys, angers, griefs, and pleasures, and discoursed on the etiquette for entertaining guests, religious rites, and the principles of being a wise mother and a good wife. *Kasa* written by the commoners were marked by the same style as those written by women and played a similar role in the literary activity of the general masses. There are two other forms of *kasa* written by the literati: travel records and accounts of life in exile. To the first belong *Iltong changyu ka* (1764; "Song of a Grand Trip to Japan"), written by Kim In'gyŏm upon his return from an official trip to Japan; and the *Yŏnhaeng ka* (1866; "Song of a Journey to Peking"), written by Hong Sunhak upon his return from an official trip to Peking. The second includes *Pukkwan kok* ("Song of the Northern Pass"), written by Song Chusŏk, who in 1675 accompanied his grandfather, Song Siyŏl, to his place of exile in the northeast; the *Manŏnsa* ("Song of Ten Thousand Words"), written by An Towŏn (or An Chohwan) during his banishment on the lonely island of Ch'uja, off the southeast coast of Korea; and the *Pukch'ŏn ka* (1853; "Song of a Northern Exile"), written by Kim Chinhyŏng, depicting the life of exile in the northeast.

Oral literature. Another feature of the later Yi dynasty was the formation, by the common people, of *p'ansori* texts. *P'ansori* seem to have originated during the reign of Sukchong (1675–1720), when old folktales were first sung. Their style and form were fixed by the *kwangdae,* or professional singers, and a group of amateurs in Chŏlla and Ch'ungch'ŏng provinces. Six of the original 12 titles were revised by the master *p'ansori* writer Sin Chaehyo, of which five are still performed.

Mask plays and puppet shows

The representative mask play is *Sandae kŭk.* Of unknown origin, it was usually performed on a makeshift, open-air stage in 12 scenes, or acts. The masked actors followed a script that presented a story in dialogue interspersed with dances and songs. As the puppets of the *Kkoktukaksi nori* show were made of *pak* (a gourd, rhyming with the Korean name Pak), it was also called *Pak Ch'omji kŭk* ("Old Pak's Play"). Through keen satire presented in a unique and distinguished style, the contents of the masked drama and the puppet show strongly reflect the environment and the feelings of the common people of the later Yi dynasty.

Transitional literature: 1894–1910. By the time of the 1894 reforms, enough social and intellectual change had occurred to suggest the beginnings of a division between traditional and modern literature. But, just as conservatism did not favour sudden changes in the political and social structure, literature, too, faced a period of transition toward its modern transformation. Schools were established by the educational ordinance of 1895, and the organization of learned societies and "enlightenment" movements followed soon after. Vernacular publications, the *Tongnip sinmun* ("Independent") and the *Cheguk sinmun* ("Imperial Post"), along with the establishment of the Korean Language Institute and the scientific study, consolidation, and systematization of Korean grammar, also helped open the way for the modern literary movement.

The first literary forms to appear after the 1894 reforms were the *sinsosŏl* ("new novel") and the *ch'angga* ("song"). These transitional literary forms were stimulated by the adaptation of foreign literary works and the rewriting of traditional stories in the vernacular. The *ch'angga,* which evolved from hymns sung at churches and schools in the 1890s, became popular upon the publication of the "Aeguk ka" ("National Anthem"), by Yi Yongu, and "Tongsim ka" ("A Boy's Mind"), by Yi Chungwŏn, in an issue (1896) of the *Tongnip sinmun.* Songwriters still used such traditional verse forms as the *sijo* and *kasa* or a song form, the predominant pattern of which (seven and five syllables) showed the influence of popular Japanese songs (*shōka*). Most songs denounced corruption in the government and stressed independence, patriotic fervour, and modernization.

The *sinsosŏl*

Three distinctly traditional elements were inherited by the *sinsosŏl.* First was the basic moral stance of reproving vice and rewarding virtue. Owing to the prevailing atmosphere of the "enlightenment" period, advocates of modernization were cast as virtuous, while the wicked were conservative. Second, the development of the plot was governed by coincidence, and events that lacked causality were nevertheless arbitrarily connected. Finally, the dialogue and the accompanying narrative were fused into one expository structure. The pioneering aspects of the *sinsosŏl,* however, were that it was written wholly in prose, whereas a considerable part of traditional fiction had been in verse; and it tried to depict a plausible human existence with backgrounds and events that more closely resembled reality than was the case in traditional fiction, which tended to follow certain model stories with their established plot lines and stereotyped characterizations. Writers of *sinsosŏl* also tried to unify the spoken and written language. Typical writers and their works are Yi Injik, *Kwi ŭi sŏng* (1907; "A Demon's Voice"); Yi Haejo, *Chayujong* (1910; "Liberty Bell"); and Ch'oe Ch'ansik, *Ch'uwŏlsaek* (1912; "Colour of the Autumn Moon"). In their works these writers advocated modernization, a spirit of independence, contact with advanced countries, study abroad, the diffusion of science and technology, and the abolition of conventions and superstition.

Modern literature: 1910 to the present. The modern literary movement was launched by Ch'oe Namsŏn and Yi Kwangsu. In 1908 Ch'oe published the poem "Hae egeso pada ege" ("From the Sea to Children") in *Sonyŏn* ("Children"), the first literary journal aimed at producing cultural reform. Inspired by Byron's *Childe Harold's Pilgrimage,* Ch'oe celebrates, in clean masculine diction, the strength of the young people who will carry out the necessary social and literary revolution. The poem's inventions include the use of punctuation marks, stanzas of unequal length, and reference to the sea and children, hitherto little mentioned in classical poetry. Neither Ch'oe nor his contemporaries, however, could escape the bounds of traditional prosody or succeed in modernizing traditional forms of speech and allusion. In his stories, which dealt with the enlightened pioneers who championed Western science and civilization, Yi Kwangsu adopted a prose style that approximated the everyday speech of common people. Yi's reputation was established by *Mujŏng* (1917; "The Heartless"), the first modern Korean novel.

In 1919, shortly before the unsuccessful movement for independence from Japan, translations of such Western poets as Paul Verlaine, Rémy de Gourmont, and Stéphane Mallarmé began to exert a powerful influence on Korean poetry. The indirection and suggestiveness of French Symbolist literature were introduced by Kim Ŏk, the principal translator. Against the didacticism of the age Kim set Mallarmé, and against its rhetoric and sentimentality he set Verlaine, concluding in the process that free verse was the supreme creation of the Symbolists. Kim's fascination with the Symbolist movement culminated in the publication of *Onoe ŭi mudo* (1921; "Dance of Anguish"), the first Korean collection of translations from Western poetry. The exotic and melancholy beauty of autumn and expressions of ennui and anguish appealed to poets who sought to vent their frustration and despair at the collapse of the independence movement.

Naturalism

The movement for literary naturalism was launched in the 1920s by a group of young writers who rallied around a new definition of universal reality. Yŏm Sangsŏp, the first to introduce psychological analysis and scientific documentation into his stories, defined naturalism as an expression of awakened individuality. Naturalism's purpose, Yŏm asserted, was to expose the sordid aspects of reality, especially the sorrow and disillusionment occurring as authority figures are debased and one's idols are shattered. Many works of naturalist fiction were first-person narratives in which writers presented themselves as the subjects of case studies. The disharmony between the writer and his society often induced the writer to turn to nature; the land and simple folk furnished themes and motifs for some of the better stories in the Zolaesque tradition, among them "Pul" (1925; "Fire") by Hyŏn Chingŏn and "Kamja" (1925; "Potato") by Kim Tongin.

The 1920s produced several major poets, Han Yongun published *Nim ŭi ch'immuk* (1926; "The Silence of Love"), comprising 88 meditative poems. Han sought insight into the reasons why he and his country had to endure Japanese occupation, and he found Buddhist contemplative poetry the lyric genre most congenial to this pursuit. The nature and folk poet Kim Sowŏl used simplicity, directness, and terse phrasing to good effect. Many of his poems in *Chindallaekkot* (1925; "Azaleas") were set to music.

The Mukden, or Manchurian, Incident (1931) and the Japanese invasion of China in 1937 induced the Japanese military authorities to impose wartime restrictions. The grinding poverty of the lower classes at home and abroad, especially in the Korean settlements in southern Manchuria, was the chief concern of the writers of the "new tendency" movement, which opposed the romantic and "decadent" writers of the day and later became proletarian in spirit. Writers of the class-conscious Korean Artist Proletariat Federation (KAPF), organized in 1925, asserted the importance of propaganda and regarded literature as a means to establish socialism.

Modern Korean literature attained its maturity in the 1930s through the efforts of a group of talented writers. They drew freely upon European examples to enrich their art. Translation of Western literature continued, and works by I.A. Richards, T.S. Eliot, and T.E. Hulme were introduced. This artistic and critical activity was a protest against the reduction of literature to journalism and its use as propaganda by leftist writers.

The first truly successful poet of modern Korea was Chŏng Chiyong, who was influenced by William Blake and Walt Whitman. *Paengnoktam* (1941; "White Deer Lake"), his second book of poetry, symbolically represents the progress of the spirit to lucidity and the fusion of man and nature. A poetry of resistance, voicing sorrow for the ruined nation with defiance but without violence or hatred, was produced by Yi Yuksa and Yun Tongju. In Yi's poem "Chŏlchŏng" (1939; "The Summit"), he recreates the conditions of an existence in extremity and forces the reader to contemplate his ultimate destiny. The poetry of Yun Tongju, a dispassionate witness to Korea's national humiliation, expresses sorrow in response to relentless tyranny.

Korean fiction of the 1930s took shape in the void created by the compulsory dissolution of KAPF in 1935. Barred from all involvement with social or political issues, some writers returned to nature and sex; others retreated to the labyrinth of primitive mysticism, superstition, and shamanism; still others sympathetically portrayed characters born out of their time, defeated and lonely. In the early 1940s, the Japanese suppressed all writings in Korean. Censorship, which had begun with the Japanese annexation of Korea in 1910, was intensified. Korea was liberated in August 1945, and the Republic of Korea (South Korea) was established three years later. The literary scene experienced the revival of the controversy between left and right that had raged in the late 1920s and early 1930s. There were frantic groupings and regroupings, and most of the hardcore leftist writers, such as Yi Kiyŏng and Han Sŏrya, were in North Korea by 1948.

Literature since World War II

The liberation of 1945 produced a flowering of poetry of all kinds. Some poets were determined to bear witness to the events of their age; some sought to further assimilate traditional Korean values, while others drew variously on Western traditions to enrich their work. Sŏ Chŏngju and Pak Tujin are known for their lifelong dedication and contributions to modern Korean poetry. Considered to be the most "Korean" of contemporary poets, Sŏ is credited with exploring the hidden resources of the language, from sensual ecstasy to spiritual quest, from haunting lyricism to colloquial earthiness. Pak is capable of a wide range of moods, and his language and style impart a distinctive tone to his Christian and nationalistic sentiments. Marked by sonorific intricacies and incantatory rhythms, Pak's poems are imbued with a strong historical and cultural consciousness that bears testimony to contemporary reality.

The single overwhelming reality in Korean fiction since the Korean War has been the division of the country. The 38th parallel torments the conscience of every fictional protagonist, for it is a symbol not only of Korea's trials but also of the division of mankind and of the protagonist's alienation from himself and his world. Some have attempted to capture the images of the people in lyrical prose; others have delved into the conscience of the war's lost generation or into the inaction, self-deception, and boredom of the alienated generation of the 1960s. Some have studied the defeat and disintegration of good people; others have investigated the ways in which modern society negates freedom and individuality. Outstanding among writers of the roman-fleuve is Pak Kyŏngni, the mother-in-law of the poet Kim Chiha. Pak's multivolume *T'oji* (1969; "Land") has been acclaimed for its commanding style and narrative techniques.

In the last quarter of the 20th century a host of talented writers have been perfecting the art of being themselves. The poet Hwang Tonggyu, for example, has drawn material not only from his own experiences but also from the common predicament of the Korean people, expressing what others know but do not think of saying or cannot say. The novelist Yun Hŭnggil is another example of a writer who has cultivated fiction as an instrument of understanding himself and others. In his *Changma* (1973; "The Rainy Spell"), for example, Yun says that ideological differences imposed upon the Korean people by history can be overcome if they delve into the native traditions that have given them cohesion.

Drama

The "new" drama movement, which began in 1908, saw the rise and fall of small theatre groups, such as the T'owŏrhoe, organized in 1923, and finally the Kŭk Yesul Yŏnguhoe ("Theatrical Arts Research Society"), organized in 1931. Through their experimental theatre, the members of the society staged contemporary Western plays and encouraged the writing of original plays, such as Yu Ch'ijin's *T'omak* (1933; "Clay Hut"). The paucity of first-rate playwrights and actors, the dearth of plays that satisfy dramatic possibilities, and the general living standards of the audience, as well as the lack of government support, have limited the scope of dramatic activity. Domestic plays and historical pieces, however, have continued to be written and staged.

BIBLIOGRAPHY

Poetry: PETER H. LEE (comp. and ed.), *Anthology of Korean Literature from Early Times to the Nineteenth Century* (1981), collects representative poetic and prose works written in Chinese and Korean and supplies commentary and criticism; his *Lives of Eminent Korean Monks* (1969) is an annotated translation of KAKHUN, *Haedong kosŭng chŏn* (1215), with an introduction. RICHARD RUTT (ed. and trans.), *The Bamboo Grove* (1971), introduces *sijo* arranged by themes. WON KO (trans. and comp.), *Contemporary Korean Poetry* (1970), is another collection. PETER H. LEE (ed.), *The Silence of Love: Twentieth-Century Korean Poetry* (1980), contains translations of 16 major modern poets. DAVID R. MCCANN (trans.), *The Middle Hour: Selected Poems of Kim Chi Ha* (1980), contains 40 poems. Another selection of poems, prose pieces, and a play by the same author is presented in CHONG SUN KIM and SHELLY KILLEN (eds.), *The Gold Crowned Jesus and Other Writings* (1978).

Prose: RICHARD RUTT and CHONG-UN KIM (trans.), *Virtuous Women* (1974, reprinted 1979), contains translations of "Dream of Nine Clouds," "Tale of Queen Inhyŏn," and "The Song of a Faithful Wife, Ch'un Hyang." IN-SŎP CHŎNG (ed. and trans.), *Folk Tales from Korea* (1952, reprinted 1969), is a representative selection. DUK-SOON CHANG *et al.* (eds.), *The Folk Treasury of Korea: Sources in Myth, Legends, and Folktale,* trans. by TAE-SUNG KIM (1970), is a collection of oral literature. SOUN KIM, *The Story Bag* (1955), collects 30 folktales. For modern prose see KEVIN O'ROURKE (comp.), *Ten Korean Short Stories* (1973, reissued 1981); PETER H. LEE (ed.), *Flowers of Fire: Twentieth-Century Korean Stories* (1974, rev. ed. 1986); CHONG-WHA CHUNG (ed.), *Modern Korean Short Stories* (1980); CHONG-UN KIM (ed.), *Postwar Korean Short Stories,* 2nd ed. (1983); and JI-MOON SUH (trans.), *The Rainy Spell and Other Korean Stories* (1983).

Literary criticism: PETER H. LEE, *Korean Literature: Topics and Themes* (1965), is an introduction to Korean literature, and his *Songs of Flying Dragons: A Critical Reading* (1974) is an annotated translation of *Yongbi ŏch'ŏn ka* (1445–47). W.E. SKILLEND, *Kodae Sosŏl: A Survey of Korean Traditional Style Popular Novels* (1969), is a catalog of Korean fiction.

(B.-W.C./P.H.L.)

The Korean War

The Korean War was a conflict between the Democratic People's Republic of Korea (North Korea) and the Republic of Korea (South Korea) in which at least 2.5 million persons lost their lives. The war reached international proportions in June 1950 when North Korea, supplied and advised by the Soviet Union, invaded the South. The United Nations, with the United States as the principal participant, joined the war on the side of the South Koreans, and the People's Republic of China came to North Korea's aid. After more than a million combat casualties had been suffered on both sides, the war ended in July 1953 with Korea still divided into two hostile states. Negotiations in 1954 produced no further agreement, and the front line has been accepted ever since as the de facto boundary between North and South Korea.

This article is divided into the following sections:

Revolution, division, and partisan warfare, 1945–50

The Korean War had its immediate origins in the collapse of the Japanese empire at the end of World War II in September 1945. Unlike China, Manchuria, and the former Western colonies seized by Japan in 1941–42, Korea, annexed to Japan since 1910, did not have a native government or a colonial regime waiting to return after hostilities ceased. Most claimants to power were harried exiles in China, Manchuria, Japan, the U.S.S.R., and the United States. They fell into two broad categories. The first was made up of committed Marxist revolutionaries who had fought the Japanese as part of the Chinese-dominated guerrilla armies in Manchuria and China. One of these exiles was a minor but successful guerrilla leader named Kim Il-sung, who had received some training in Russia and had been made a major in the Soviet army. The other Korean nationalist movement, no less revolutionary, drew its inspiration from the best of science, education, and industrialism in Europe, Japan, and America. These "ultranationalists" were split into rival factions, one of which centred on Syngman Rhee, educated in the United States and at one time the president of a dissident Korean Provisional Government in exile.

Division at the 38th parallel

In their hurried effort to disarm the Japanese army and repatriate the Japanese population in Korea (estimated at 700,000), the United States and the Soviet Union agreed in August 1945 to divide the country for administrative purposes at the 38th parallel (latitude 38° N). At least from the American perspective, this geographical division was a temporary expedient; however, the Soviets began a short-lived reign of terror in northern Korea that quickly politicized the division by driving thousands of refugees south. The two sides could not agree on a formula that would produce a unified Korea, and in 1947 U.S. President Harry S. Truman persuaded the United Nations (UN) to assume responsibility for the country, though the U.S. military remained nominally in control of the South until 1948. Both the South Korean national police and constabulary doubled in size, providing a southern security force of about 80,000 by 1947. In the meantime, Kim Il-sung strengthened his control over the Communist Party as well as the northern administrative structure and military forces. In 1948 the North Korean military and police numbered about 100,000, reinforced by a group of southern Korean guerrillas based at Haeju in western Korea.

The creation of an independent South Korea became UN policy in early 1948. Southern communists opposed this, and by autumn partisan warfare had engulfed parts of every Korean province below the 38th parallel. The fighting expanded into a limited border war between the South's newly formed Republic of Korea Army (ROKA) and the North Korean border constabulary as well as the North's Korean People's Army (KPA). The North launched 10 cross-border guerrilla incursions in order to draw ROKA units away from their guerrilla-suppression campaign in the South.

In its larger purpose the partisan uprising failed: the Republic of Korea was formed in August 1948, with Syngman Rhee as president. Nevertheless, almost 8,000 members of the South Korean security forces and at least 30,000 other Koreans lost their lives. Many of the victims were not security forces or armed guerrillas at all but simply people identified as "rightists" or "reds" by the belligerents. Small-scale atrocities became a way of life.

The partisan war also delayed the training of the South Korean army. In early 1950, American advisers judged that fewer than half of the ROKA's infantry battalions were even marginally ready for war. American military assistance consisted largely of surplus light weapons and supplies. Indeed, General Douglas MacArthur, commander of the United States' Far East Command (FECOM), argued that his Eighth Army, consisting of four weak divisions in Japan, required more support than the Koreans.

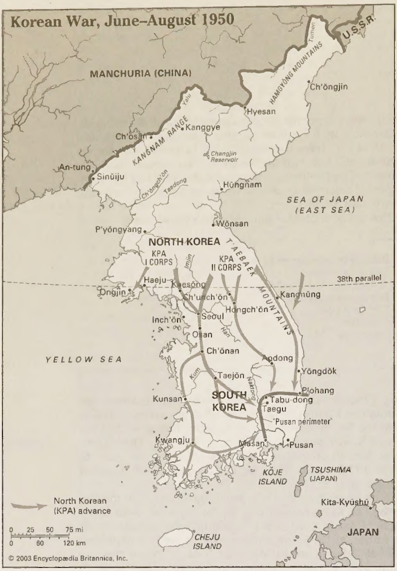
Invasion and counterinvasion, 1950–51

SOUTH TO PUSAN

In early 1949, Kim Il-sung pressed his case with Soviet leader Joseph Stalin that the time had come for a conventional invasion of the South. Stalin refused, concerned about the relative unpreparedness of the North Korean armed forces and about possible U.S. involvement. In the course of the next year, the communist leadership built the KPA into a formidable offensive force modeled after a Soviet mechanized army. The Chinese released Korean veterans from the People's Liberation Army, while the Soviets provided armaments. By 1950 the North Koreans enjoyed substantial advantages over the South in every category of equipment. After another Kim visit to Moscow in March–April 1950, Stalin approved an invasion.

The invasion launched

In the predawn hours of June 25, the North Koreans struck across the 38th parallel behind a thunderous artillery barrage. The principal offensive, conducted by the KPA I Corps (53,000 men), drove across the Imjin River toward Seoul. The II Corps (54,000 soldiers) attacked along two widely separated axes, one through the cities of Ch'unch'ŏn and Inje to Hongch'ŏn and the other down the east coast road toward Kangnŭng. (See the map, June–Au-



gust 1950.) The KPA entered Seoul in the afternoon of June 28, but the North Koreans did not accomplish their goal of a quick surrender by the Rhee government and the disintegration of the South Korean army. Instead, remnants of the Seoul-area ROKA forces formed a defensive line south of the Han River, and on the east coast road ROKA units gave ground in good order. Still, if the South was to stave off collapse, it would need help—from the U.S. armed forces.

Truman's initial response was to order MacArthur to transfer munitions to the ROKA and to use air cover to protect the evacuation of American citizens. Instead of pressing for a congressional declaration of war, which he regarded as too alarmist and time-consuming when time was of the essence, Truman went to the United Nations for sanction. Under American guidance, the UN called for the invasion to halt (June 25), then for the UN member states to provide military assistance to the Republic of Korea (June 27). By charter the Security Council considered and passed the resolutions, which could have been vetoed by a permanent member such as the Soviet Union. The Soviets, however, were boycotting the Security Council over the issue of admitting the People's Republic of China to the UN. Congressional and public opinion in the United States, meanwhile, supported military intervention without significant dissent.

Having demonstrated its political will, the Truman administration faced the unhappy truth that it did not have much effective military power to meet the invasion. MacArthur secured the commitment of three U.S. divisions from Japan, but adding ground forces only expanded the scope of defeat. For almost eight weeks near Osan, along the Kum River, through Taejŏn, and south to Taegu, American soldiers fought and died—and some fled. Weakened by inadequate weapons, limited numbers, and uncertain leadership, U.S. troops were frequently beset by streams of refugees fleeing south, which increased the threat of guerrilla infiltration. These conditions produced unfortunate attacks on Korean civilians, such as the firing on hundreds of refugees at a railroad viaduct near the hamlet of No Gun Ri, west of the Naktong River, during the last week in July.

United Nations Command fights back

It was not until the first weeks of August that the United Nations Command, or UNC, as MacArthur's theatre forces had been redesignated, started to slow the North Koreans. The Eighth Army, commanded by Lieutenant General Walton H. Walker, one of the best corps commanders in Europe in 1944–45, and the ROKA, led by Major General Chung Il-kwon, rallied and fought back with more success. Supplies came through the port at Pusan, where the Eighth Army's logistics system depended on Korean and Japanese technicians and on thousands of Korean labourers. To stop the North Koreans' tanks and supporting artillery and infantry, Walker brought in Sherman and Pershing medium tanks, rocket launchers, artillery pieces, antiaircraft guns, and, most important of all, close-air-support aircraft. The Fifth Air Force attacked forward units of the KPA with World War II-era P-51 Mustangs, new jet-powered F-80s and F-84s, and even B-26 and B-29 bombers. U.S. Marine Corps squadrons, embarked on navy light carriers, were capable of flying anywhere along the front in quick response to requests from ground forces, and on the east coast the U.S. Navy's cruisers and destroyers became a seagoing heavy artillery for the ROK I Corps. Meanwhile, fresh U.S. Army and Marine Corps units began to arrive, supplemented by a British Commonwealth brigade. In the same period, the ROKA, which had shrunk to half its prewar strength through deaths, surrenders, a few defections, and substantial desertions, began to bring its ranks back up with reservists, student volunteers, and men impressed from cities' streets as the South Koreans fell back.

Concerned that the shift of combat power toward the UNC would continue into September, the field commander of the KPA, General Kim Chaek, ordered an advance against the Naktong River–Taegu–Yŏngdŏk line, soon to become famous as the "Pusan Perimeter." The major effort was a double envelopment of Taegu, supplemented by drives toward Masan and P'ohang, the southwestern and northeastern coastal anchors of the perimeter. None reached significant objectives. At the Battle of Tabu-dong (August 18–26), the ROK 1st Division and the U.S. 27th

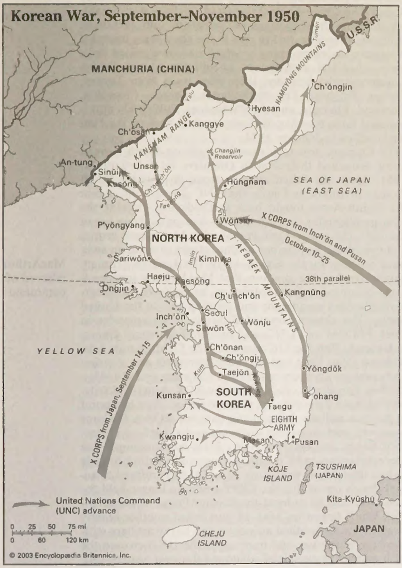
© Hulton-Deutsch Collection/Corbis

South Korean troops marching past an American tank crew near Taejŏn, 1950.

Regimental Combat Team defeated the North Koreans' main armoured thrust toward Taegu. By September 12 the KPA, its two corps reduced to 60,000 men and its tank forces destroyed, had been driven back in most places west of the Naktong and well away from Taegu and P'ohang. At that moment the entire strategic balance of the war was shifted by the sudden appearance of X Corps at Inch'ŏn.

NORTH TO THE YALU

MacArthur did not believe that he could win the war without an amphibious landing deep behind enemy lines, and he had started to think about a landing as early as July. As the core of his landing force he and the Joint Chiefs of Staff selected the 1st Marine Division and the Eighth Army's remaining infantry division, the 7th. As the force developed, it also included South Korean marine and infantry units and an assortment of U.S. support troops. The entire force was designated X Corps and was commanded by Major General Edward M. Almond, MacArthur's chief of staff.



Korean War, September–November 1950

Landing at Inch'ŏn

For the landing site, MacArthur himself fixed on Inch'ŏn, the port outlet of Seoul on Korea's west coast. A host of problems defied a landing there: wide tidal variance, mines, a crazy quilt of islands and shoal waters, and dangerous proximity to KPA reinforcements from Seoul. MacArthur brushed off all these concerns. After a naval gun and aerial bombardment on September 14, marines the next day assaulted a key harbour defense site, Wolmi Island, and then in the late afternoon took Inch'ŏn itself (see the map, September–November 1950). The North Korean resistance was stubborn but spread thin, and the 1st Marine Division, accompanied by ROK and U.S. army units, entered Seoul on September 25. The bulk of the 7th Division advanced to Suwŏn, where it contacted the Eighth Army on the 26th. MacArthur and Syngman Rhee marched into the damaged capitol building and declared South Korea liberated.

As an organized field force, the KPA disintegrated, having lost 13,000 as prisoners and 50,000 as casualties in August and September. Nevertheless, about 25,000 of its best troops took to the mountains and marched home as cohesive units; another 10,000 remained in South Korea as partisans. As the communists headed north, they took thousands of South Koreans with them as hostages and slave labourers and left additional thousands executed in their wake—most infamously at Taejŏn, where 5,000 civilians were massacred. The ROK army and national police, for their part, showed little sympathy to any southern communists they found or suspected, and American aircraft attacked people and places with little restraint. As a result, the last two weeks of September saw atrocities rivaling those seen in Europe during the fratricidal Thirty Years' War of the 17th century.

Even before the Inch'ŏn landing, MacArthur had thought ahead to a campaign into North Korea, though his plans never went beyond establishing a line across the so-called waist of Korea, from P'yŏngyang in the west to Wŏnsan in the east. On September 27 the Joint Chiefs gave him final authority to conduct operations north of the 38th parallel; however, he was instructed to limit operations in the event of Russian or Chinese intervention. For the UNC the war aim was expanded. As announced by the UN General Assembly on October 7, it was to include the occupation of all of North Korea and the elimination of the KPA as a threat to the political reconstruction of Korea as one nation. To that end, ROKA units crossed the parallel on October 1, and U.S. Army units crossed on October 7. The ROK I Corps marched rapidly up the east coast highway, winning the race for Wŏnsan; P'yŏngyang fell to the U.S. I Corps on October 19. The Kim Il-sung government, with the remnants of nine KPA divisions, fell back to the mountain town of Kanggye. Two other divisions, accompanied by Soviet advisers and air defense forces, struggled northwest toward the Yalu River and the Chinese border at Sinŭiju. The UNC assumed that the KPA had lost its will to fight. In reality, it was awaiting rescue.

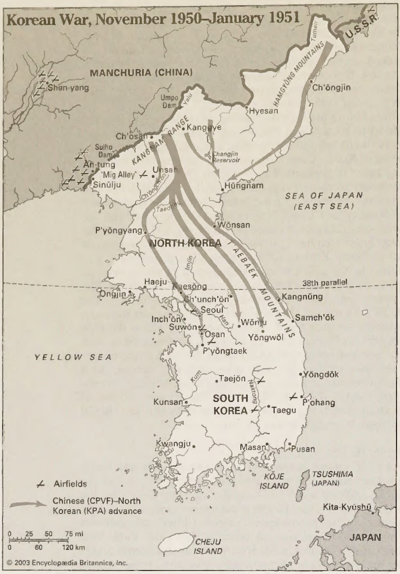
BACK TO THE 38TH PARALLEL

As UNC troops crossed the 38th parallel, Chinese Communist Party Chairman Mao Zedong received a plea for direct military aid from Kim Il-sung. The chairman was willing to intervene, but he needed assurances of Soviet air power. Stalin promised to extend China's air defenses (manned by Soviets) to a corridor above the Yalu, thus protecting air bases in Manchuria and hydroelectric plants on the river, and he also promised new Soviet weapons and armaments factories. After much debate, Mao ordered the Renmin Zhiyuanjun, or Chinese People's Volunteers Force (CPVF), to cross into Korea. It was commanded by General Peng Dehuai, a veteran of 20 years of war against the Chinese Nationalists and the Japanese.

The Chinese intervene

The Chinese First Offensive (October 25–November 6, 1950) had the limited objective of testing U.S.-ROK fighting qualities and slowing their advance. In the battle of Onjong–Unsan along the Ch'ŏngch'ŏn River, the Chinese ruined seven Korean and American regiments—including the only Korean regiment to reach the Yalu, cut off in the vastness of the cold northern hills near Ch'osan. The Chinese suffered 10,000 casualties, but they were convinced that they had found a formula for fighting UNC forces: attack at night, cut off routes of supply and withdrawal, ambush counterattacking forces, and exploit all forms of concealment and cover. Stunned by the suddenness of the Chinese onslaught and almost 8,000 casualties (6,000 of them Koreans), the Eighth Army fell back to the south bank of the Ch'ŏngch'ŏn and tightened its overextended lines. With a harsh winter beginning and supplies in shortage, the pause was wise.

Another matter of concern to the UNC was the appearance of MiG-15 jet fighters above North Korea. Flown by Soviet pilots masquerading as Chinese and Koreans, the MiGs, in one week's action (November 1–7), stopped most of the daytime raids on North Korea. The U.S. Air Force immediately dispatched a crack wing of F-86 Sabrejet interceptors to Japan, and thus a two-and-a-half year battle for air superiority was joined. Over the course of the war, the F-86s succeeded in allowing the Far East Air Forces (FEAF) to conduct offensive air operations anywhere in North Korea, and they also protected the Eighth Army from communist air attack. However, they were never able to provide perfect protection for B-29s flying daylight raids into "MiG Alley," a corridor in northwestern Korea where MiGs based near An-tung, Manchuria (now Tan-tung, China), fiercely defended bridges and



dams on the Yalu River. (See the map, November 1950–January 1951.)

The FEAF also turned its fury on all standing structures that might shield the Chinese from the cold; cities and towns all over North Korea went up in flames. But the air assault did not halt the buildup for the Chinese Second Offensive. This time Peng's instructions to his army commanders stressed the necessity to lure the Americans and "puppet troops" out of their defensive positions between the Ch'ŏngch'ŏn and P'yŏngyang, giving the impression of weakness and confusion, while Peng would surround their forward elements with his much-enlarged force of 420,000 Chinese and North Korean regulars. MacArthur, in what may have been his only real military mistake of the war, ordered the Eighth Army and X Corps northward into the trap on November 24, and from November 25 to December 14 the Chinese battered them back to South Korea. Falling upon the U.S. IX Corps and the ROK II Corps from the east, Peng's Thirteenth Army Group opened up a gap to the west and almost cut off the I Corps north of the Ch'ŏngch'ŏn. The I Corps managed to fight its way through Chinese ambushes back to P'yŏngyang. In the eastern sector the Chinese Ninth Army Group sent two armies against the 1st Marine Division near the Changjin Reservoir (known to the Americans by its Japanese name, Chosin). Under the worst-possible weather conditions, the marines turned and fought their way south, destroying seven Chinese divisions, before reaching sanctuary at the port of Hŭngnam on December 11.

U.S. Marines at Chosin

At the height of the crisis, MacArthur conferred with Walker and Almond, and they agreed that their forces would try to establish enclaves in North Korea, thus preserving the option of holding the P'yŏngyang–Wŏnsan line. In reality, Walker had finally reached the limits of his disgust with MacArthur's meddling and posturing, and he started his men south. By December 6 the Eighth Army had destroyed everything it could not carry and had taken the road for Seoul. Walker's initiative may have saved his army, but it also meant that much of the rest of the war would be fought as a UNC effort to recapture ground surrendered with little effort in December 1950. Walker himself died in a traffic accident just north of Seoul on December 23 and was succeeded by Lieutenant General Matthew B. Ridgway.

Heartened by the ease with which the CPVF had driven the UNC out of North Korea, Mao Zedong expanded his war aims to demand that the Chinese army unify all of Korea and drive the Americans and puppets off the peninsula. His enthusiasm increased when the Chinese Third Offensive (December 31, 1950–January 5, 1951) retook Seoul. The Chinese attacks centred on ROKA divisions, which were showing signs of defeatism and ineptness. Ridgway, therefore, had to rely in the short term upon his American divisions, many of which had now gained units from other UN participants. In addition to two British Commonwealth brigades, there were units from Turkey, France, Belgium, The Netherlands, Greece, Colombia, Thailand, Ethiopia, and the Philippines. Pulling his multinational force together, Ridgway pushed back to the Han River valley in January 1951.

The Chinese, now reinforced by a reborn North Korean army, launched their Fourth Offensive on February 11, 1951. Again the initial attacks struck ill-prepared South Korean divisions, and again the UNC gave ground. Again the Eighth Army fought back methodically, crossing the 38th parallel after two months. At that point Peng began the Fifth Offensive (First Phase) with 11 Chinese armies and two North Korean corps. The attacks came at an awkward moment for the Eighth Army. On April 11 Truman, having reached the opinion that MacArthur's independence amounted to insubordination, had relieved the general of all of his commands and recalled him to the United States. The change elevated Ridgway to commander in chief, FECOM and UNC, and brought Lieutenant General James A. Van Fleet to command the Eighth Army. Like Ridgway, Van Fleet had earned wide respect as a division and corps commander against the Germans in 1944–45.

MacArthur relieved of command

Before Van Fleet could reform the ROK Army and redeploy his own divisions, the Chinese struck. At a low point in Korean military history, the battered ROKA II Corps gave way, and American divisions peeled back to protect their flanks and rear until Van Fleet could commit five more American and Korean divisions and a British brigade to halt the Chinese armies on April 28. Mao refused to accept Peng's report that the CPVF could no longer hold the initiative, and he ordered the Second Phase of the offensive, which began on May 16 and lasted another bloody week. Once again allied air power and heavy artillery stiffened the resistance, and once again the UNC crossed the 38th parallel in pursuit of a battered (but not beaten) Chinese expeditionary force.

TO THE NEGOTIATING TABLE

By June 1951 the Korean War had reached another critical point. The Chinese–North Korean armies, despite having suffered some 500,000 casualties since November, had grown to 1,200,000 soldiers. United Nations Command had taken its share of casualties—more than 100,000 since the Chinese intervention—but by May 1951 American ground troops numbered 256,000, the ROKA 500,000, and other allied contingents 28,000. The U.S. Far East Air Forces had grown from fewer than 700 aircraft in July 1950 to more than 1,400 in February 1951.

These developments obliged the leaders of both coalitions to consider that peace could not be imposed by either side through military victory—at least at acceptable cost. Truman and the UN, in particular, had lost their ardour for anything more than a return to status quo ante bellum and were sympathetic to the idea of a negotiated settlement. On May 17, 1951, the U.S. National Security Council adopted a new policy that committed the United States to support a unified, democratic Korea, but not necessarily one unified by military action and the overthrow of Kim Il-sung.

The communist road to a negotiated peace started in Peking, where Mao, who had no desire to end the war, approved an approach suggested by Peng and others: hold the ground in Korea and conduct a campaign of attrition, attempting to win limited victories against small allied units through violent night attacks and infantry infiltration. Protection from UNC aircraft and artillery would be provided by caves and bunkers dug into the Korean mountains. Meanwhile, negotiations would be managed by the Chinese, an unparalleled chance to appear an equal of the United States in Asia and a slap at the hated Japanese. The Koreans were not a factor for either side.

Agreement to meet

After secret meetings between U.S. and Soviet diplomats, the Soviet Union announced that it would not block a negotiated settlement to the Korean War. The Truman administration had already alerted Ridgway to the prospect of truce talks, and on June 30 he issued a public statement that he had been authorized to participate in "a meeting to discuss an armistice providing for the cessation of hostilities." On July 2 the Chinese and North Koreans issued a joint statement that they would discuss arrangements for a meeting, but only at their place of choice: the city of Kaesŏng, an ancient Korean capital, once part of the Republic of Korea but now occupied by the communists at the very edge of the front lines. The Chinese had just fired the first salvo of a new war, one in which talking and fighting for advantage might someday end the conflict.

Talking and fighting, 1951–53

BATTLING FOR POSITION

From the time the liaison officers of both coalitions met on July 8, 1951, until the armistice agreement was signed on July 27, 1953, the Korean War continued as a "stalemate." This characterization is appropriate in only two ways: (1) both sides had given up trying to unify Korea by force; and (2) the movement of armies on the ground never again matched the fluidity of the war's first year. Otherwise, the word stalemate has no meaning, for the political-geographic stakes in Korea remained high.

As the negotiations at Kaesŏng developed, neither Ridgway nor Van Fleet believed that the talks would produce anything without more UNC offensives beyond the 38th parallel. Ridgway was particularly convinced that UNC forces should take the "Iron Triangle" (see the map, February 1951–July 1953), a key area between the headwaters of the Imjin River and the highest eastern mountain ranges that was anchored on the cities of Chorwon (west), Pyŏnggang (north), and Kimhwa (east). Communist planners were equally convinced that control of this terrain offered advantages for defending North Korea or for continuing the war with offensives to the south and east.

Ground actions never actually ceased in 1951, but none matched the ferocity and frustration of the Eighth Army's Autumn Offensive (August 31–November 12). Van Fleet's general concept envisioned operations by the I Corps (five divisions) in the west and X Corps (five divisions) in the central-eastern sector. In the I Corps sector, the ROK 1st Division and the British Commonwealth Division made notable advances beyond the Imjin valley, while other U.S. and ROK divisions advanced past Chorwon and then stalled in heavy fighting. The X Corps, fighting a crack Chinese army and two North Korean corps, pushed northward through the mountains and succeeded only in making "Bloody Ridge," "Heartbreak Ridge," "The Punchbowl," and Kanmubong Ridge bad memories for thousands of army and marine veterans. The KPA I, III, and VI Corps, holding the eastern mountains, proved especially difficult to dislodge, for Kim Il-sung had issued a "stand or die" order to his much-enlarged and improved armed forces. The most surprising advance occurred in the X Corps sector, where two U.S. and two ROK divisions pushed the Chinese back almost 10 miles (15 kilometres) from Kimhwa to Kŭmsong, pushing the front line out in a salient that exposed their flanks but also establishing a strong position to advance west to Pyŏnggang. The cost of the campaign troubled Van Fleet and Ridgway: 60,000 casualties, 22,000 of them American.

"Bloody Ridge" and "Heartbreak Ridge"

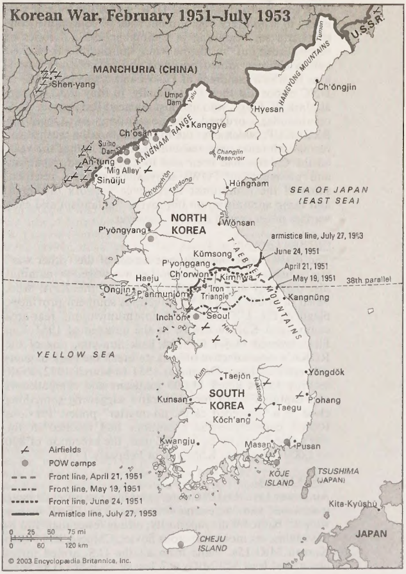
The campaign did not discourage the Chinese leadership, since in their eyes the strategy of "active defense" had worked. The UNC gave up major offensive operations in November, and the Chinese actually struck back in places with some success. Communist losses of some 100,000–150,000 were significant but not crippling—certainly not enough to drive the Chinese to end the war, only to talk some more about it.

In late October 1951 the communists agreed to move the truce negotiations to a more secure area, a village named P'anmunjŏm. Within two months they accepted the current line of contact between the armies as the military demarcation line; they also accepted related measures for the creation of a demilitarized zone (DMZ). The UNC accepted that there would be no verification activities outside of the DMZ, and both sides agreed to work on a regime for enforcement of the armistice after the shooting stopped. Much work on these items remained to be done, but the outline of an agreement was becoming apparent as the year ended—with one major exception: the handling of prisoners of war (POWs).

BATTLING OVER PRISONERS OF WAR

As another bitterly cold Korean winter congealed operations on the ground, repatriation of POWs became the most intractable issue at P'anmunjŏm. The initial assumption by the negotiators was that they would follow the revised Geneva Convention of 1949, which required any "detaining authority" that held POWs to return all of them to their homelands as rapidly as possible when a war ended. This "all for all" policy of a complete—even forced—exchange of prisoners was certainly favoured by the American military, which was alarmed by early reports from Korea of atrocities against allied POWs. The South Korean government, on the other hand, was adamantly opposed to complete and involuntary repatriation, since it knew that thousands of detainees in the South were actually South Korean citizens who had been forced to fight with the KPA. Indeed, the North Koreans knew that they had much to answer for regarding their impressment, murder, and kidnapping of South Koreans. The Chinese army

Voluntary versus involuntary repatriation



© 2003 Encyclopædia Britannica, Inc.

leaders, meanwhile, knew that some of their soldiers, impressed from the ranks of the Nationalist army, would refuse repatriation if it was not made mandatory.

Both sides agreed to exchange the names of POWs and the numbers held in various categories. The results of the tally shocked all the participants. The U.S. armed forces were carrying 11,500 men as missing in action (MIA), but the communists reported only 3,198 Americans in their custody (as well as 1,219 other UN POWs, mostly Britons and Turks). The accounting for the South Koreans was even worse: of an estimated 88,000 MIAs, only 7,142 names were listed. The numbers fed the fears of the allies that the murder of POWs had been even worse than they suspected. In truth, most of the MIAs had died in battle, but perhaps 15,000 (all but 2,000 of them South Koreans) had died in communist hands from torture, execution, starvation, and medical mistreatment.

The communists, too, found little comfort in the numbers. Early unofficial estimates of POWs in UNC custody had been either too low, around 90,000, or too high, around 170,000. Now the official list produced 95,531 North Koreans, 20,700 Chinese, and 16,243 South Koreans, for a total of 132,474. The UNC reported that the 40,000 "missing" men were South Koreans who had already passed loyalty investigations and would not be counted as potential repatriates. Against this background, Truman ruled in January 1952 that no POW in UNC custody would be forced to return to North Korea or China against his will. Koreans choosing to go north would be exchanged on a "one for one" formula until all 12,000 allied POWs had been returned. Such a process, however, would require extensive screening of individuals about their preferences, a condition that soon created open warfare in the camps.

Uprisings in the camps

The communists had taken steps in 1951 to infiltrate political officers into the UNC POW camps, and now orders came from P'yŏngyang to obstruct the screening process without regard for loss of life. The goal was to make the POWs so obnoxious that the UNC would use force if necessary to send every one of them back to communist control. And so, beginning in December 1951, a series of revolts broke out "inside the wire," culminating in pitched battles between armed prisoners and entire guard battalions in which hundreds of POWs and a small number of UNC troops lost their lives. Finally, in May 1952, General Mark W. Clark, who had just replaced Ridgway as UNC commander, ordered the execution of Operation BREAKUP, which over the following months crushed the revolt with tanks, gas, and bullets. By the end of the year, all the Chinese had been sent to Cheju Island, repatriate and nonrepatriate POWs segregated, refugees resettled, some of the communist intelligence network disrupted, and camp administration improved. Vigilantism and gang warfare never ceased entirely, however.

GUERRILLA WARFARE

The POW revolt was only one aspect of the "other war" raging behind UNC lines. Another was waged by communist partisans and stay-behind units of the KPA, who, based in South Korea's mountainous southern provinces, plagued the UNC lines of communication, rear-area camps, and Korean towns. In the autumn of 1951 Van Fleet ordered Major General Paik Sun-yup, one of the ROKA's most effective officers, to break the back of guerrilla activity. From December 1951 to March 1952, ROK security forces killed 11,090 partisans and sympathizers and captured 9,916 more—a ratio suggesting something close to a "scorched earth, no-quarter" policy. Previous ROKA counterguerrilla operations had resulted in the war's worst atrocity by a UNC unit, the execution of 800 to 1,000 villagers at Kŏch'ang in February 1951.

AIR WARFARE

Air power gave the UNC its greatest hope to offset Chinese manpower and increasing firepower. The FEAF clearly won the battle for air superiority, pitting fewer than 100 F-86s against far more numerous Soviet, Chinese, and North Korean MiG-15s. Pilots from all the U.S. armed forces downed at least 500 MiGs at a loss of 78 F-86s. The Soviets rotated squadrons of their air defense force to Korea, losing more than 200 pilots.

Strategic bombing was at first limited by policy to attacks on North Korean cities and military installations—a campaign pursued until P'yŏngyang resembled Hiroshima or Tokyo in 1945. In 1952 the bombing of power plants and dams along the Yalu was authorized, and the following year approval was given to attack dams and supporting irrigation systems in North Korea. The bombing caused great suffering for the North Koreans, but they had to follow the Chinese and Russians in the war's strategic direction, and the Chinese and Russians were hurt very little.

Considering nuclear weapons

Throughout the war U.S. political and military leaders studied the possible use of nuclear weapons, and upon four separate occasions they gave this study serious attention. The answer was always the same: existing atomic bombs, carried by modified B-29s, would have little effect except for leveling cities. The one time that Truman suggested (in December 1950) that he was considering the nuclear option, the British led the allied charge to stop such talk.

Without question, the UNC air campaign hurt the communists, and in retaliation the Chinese and North Koreans (with Soviet collusion) treated captured pilots with special brutality. Air crewmen made up the largest single group of U.S. POWs who truly disappeared, presumably dying under interrogation in Manchuria, elsewhere in China, and possibly Russia. The communists also claimed that FEAF bombers were spreading epidemic diseases among the civilian population, and they tortured captured American pilots until they extracted incriminating statements of terror bombing and germ warfare.

STRENGTHENING THE ROK

American air power might have held the communists at bay in the near term, but the long-term security of the Republic of Korea depended on (1) the enlargement and improvement of its own armed forces and (2) the stability of its government. The first requirement was accomplished by the United States' Korean Military Advisory Group, which modernized the ROKA and also organized an effective training program. In the political arena, however, the UNC had to deal with the aging Syngman Rhee, who was convinced that he had an unfinished divine mission to save Korea. In 1952 Rhee forced the National Assembly to make the election of the president a matter of popular vote, immediately calling an election and winning a second term with 5 million of the 6 million votes cast. Rhee's political coup had a ripple effect that spread to the armistice negotiations, as his dogmatic opposition to a cease-fire increased in scope and vigour. Essentially, Rhee could not believe that a likely new Republican administration in Washington, led by two other venerable Cold Warriors, Dwight D. Eisenhower and John Foster Dulles, would be satisfied to have American soldiers "die for a tie." Neither could the Russians, Chinese, and North Koreans.

THE FINAL PUSH

From September to November 1952, the Chinese expeditionary force staged its sixth major offensive of the war, this time to force the allies back to the 38th parallel and to inflict unacceptable casualties on them. Raging from the valley of the Imjin through the Iron Triangle to the eastern mountains, the ground war followed the same dismal pattern. The Chinese infiltrated allied outposts at night, then attacked under the support of short, intense artillery barrages. Submachine guns and hand grenades ruled the trenches, and flamethrowers and demolitions became standard weapons for assault units. Obscure hills acquired memorable names: White Horse Mountain, Bunker Hill, Old Baldy, Sniper Ridge, Capitol Hill, Triangle Hill, Pike's Peak, Jackson Heights, and Jane Russell Hill. By the time fighting faded in mid-November, the Eighth Army had lost 10,000 men, the Chinese 15,000. Chinese commanders hoped that they had persuaded President-elect Eisenhower to abandon any ambitious plans for a major offensive in 1953.

Waning U.S. support

The Chinese need not have worried, for both Eisenhower and Secretary of State-designate Dulles viewed continuation of the Korean War as incompatible with American national security interests. In their view the People's Republic of China was indeed the enemy in Asia, but Korea was only one theatre in the struggle. They also knew that the voting public's support for the war had thinned throughout 1952 as the talking and fighting continued abroad and the talking and taxing continued at home. As for the negotiations, Dulles conceded the communists' point that voluntary repatriation should involve screening by an international agency, not just U.S.-ROK teams. When the UN and the International Committee of the Red Cross called for an exchange of sick and disabled POWs as a goodwill gesture, Eisenhower approved.

The plan proved a good test of communist intentions—by sheer chance. On March 5, 1953, Joseph Stalin died, and within weeks the Politburo of the Soviet Communist Party voted that the war in Korea should be ended. Mao Zedong received the news with dismay, but he knew that his army could not continue the war without Soviet assistance. With a speed that amazed the negotiating teams on both sides, the Chinese accepted voluntary repatriation. POWs who wanted to return to their homelands would be released immediately, and those who chose to stay would go into the custody of a neutral international agency for noncoercive screening. The Chinese and North Koreans also agreed to the exchange of sick and disabled POWs, which took place between April 20 and May 3.

Peace was not yet at hand, however. Rhee had never publicly surrendered his "march north and unify" position, and in private he hinted that he might "accept" an armistice only in return for serious commitments by the United States, including an unambiguous mutual security alliance and $1 billion in economic aid. The Chinese, meanwhile, saw but one way to win concessions and territory in a peace agreement: on the battlefield. Their seventh and final offensive opened in the Imjin River sector in May against U.S. and Commonwealth divisions, then shifted to the South Koreans, who were driven back 20 miles (30 kilometres) from the Kŭmsong salient.

ARMISTICE

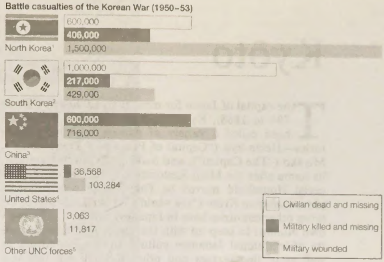
The battle of the Kŭmsong salient ended the shooting war. On May 25 the P'anmunjŏm negotiators had worked out the details of the POW exchange, making provisions for "neutral nation" management of the repatriation process. They began to plan for an armistice signing. Then on June 18–19 Syngman Rhee arranged for his military police to allow 27,000 Korean internees in their custody to "escape." Enraged, the Chinese ordered further attacks on the ROKA. The Americans shared their fury but, in the interest of compromise, convinced Rhee that the United States would meet all his preconditions for an armistice. On July 9 Rhee agreed to accept the armistice, though no representative of the ROK ever signed it. On July 27 Mark W. Clark for the UNC, Peng Dehuai for the Chinese, and Kim Il-sung for the North Koreans signed the agreement. That same day the shooting stopped (more or less), and the armies began the awkward process of disengagement across what became a 2.5-mile- (4-kilometre-) wide DMZ.

Shooting stops at the DMZ

Supervision of the armistice actions fell to a Military Armistice Commission (10 officers representing the belligerents), a Neutral Nations Supervisory Commission (Sweden, Switzerland, Poland, and Czechoslovakia), and a Neutral Nations Repatriation Commission (the same four states, plus India as the custodian of the POWs). From August 5 to September 6, 75,823 communist soldiers and civilians (all but 5,640 of them Koreans) returned to their most-favoured regime, and 7,862 ROK soldiers, 3,597 American servicemen, and 1,377 persons of other nationalities (including some civilians) returned to UNC control. The swap became a media event of potent possibilities: the communist POWs stripped off their hated capitalist prison uniforms and marched off singing party-approved songs.

The handling of those who refused repatriation turned into a nightmare, as agents among the communist POWs and interrogators made life miserable for the Indians. By the time the Neutral Nations Repatriation Commission gave up the screening process in February 1954, only 628

Battle casualties of the Korean War (1950–53)



[1] Figures reflect the higher end of U.S. and South Korean estimates, which range from 500,000 to 600,000 civilian dead and missing, from 294,000 to 406,000 military killed and missing, and from 226,000 to 1,500,000 military wounded.
[2] South Korean estimates of civilian dead and missing range from 500,000 to 1,000,000. Official figure of military killed in action is 187,712.
[3] Figures are U.S. and South Korean estimates. Official Chinese figures acknowledge 152,400 military killed and missing, 238,000 military wounded.
[4] Figures reflect official U.S. tally, which counts 33,741 battlefield killed and missing and 2,827 dead and missing in war zone from other causes. U.S. military deaths outside the war zone were 17,678, bringing the total military dead and missing during the Korean War to 54,246.
[5] No UNC member had more than 700 battlefield deaths except Turkey (721).

© 2003 Encyclopædia Britannica, Inc.

Chinese and Koreans had changed their minds and gone north, and 21,839 had returned to UNC control. Most of the nonrepatriates were eventually settled in South Korea and Taiwan.

As provided for in the armistice agreement, the United States organized an international conference in Geneva for all the belligerents to discuss the political future of Korea. The actual meetings produced no agreement. The Korean peninsula would continue to be caught in the coils of Cold War rivalry, but the survival of the Republic of Korea kept alive the hope of civil liberties, democracy, economic development, and eventual unification—even if their fulfillment might require another 50 years or more.

BIBLIOGRAPHY

Monographs. Histories of the Korean War written for a popular audience should be used with care. The best are MICHAEL HICKEY, *The Korean War: The West Confronts Communism, 1950–1953* (1999); JON HALLIDAY and BRUCE CUMINGS, *Korea: The Unknown War* (1988); CALLUM A. MACDONALD, *Korea: The War before Vietnam* (1986); KIM CHUM-KON (CHŎM-GON KIM), *The Korean War, 1950–53* (1980); and DAVID REES, *Korea: The Limited War* (1964, reissued 1970).

For students who want more detail on ground operations, the best lines of departure are ROY E. APPLEMAN, *Disaster in Korea: The Chinese Confront MacArthur* (1989), and *Ridgway Duels for Korea* (1990); CLAY BLAIR, *The Forgotten War: America in Korea, 1950–1953* (1987); and MAX HASTINGS, *The Korean War* (1987, reissued 2000).

Russian and Chinese sources, though still limited in availability, have allowed the publication of books written from the perspective of the Soviets, North Koreans, and communist Chinese. They include SHU GUANG ZHANG, *Mao's Military Romanticism: China and the Korean War, 1950–1953* (1995); CHEN JIAN (JIAN CHEN), *China's Road to the Korean War: The Making of the Sino-American Confrontation* (1994); SERGEI N. GONCHAROV, JOHN W. LEWIS, and XUE LITAI (LITAI XUE), *Uncertain Partners: Stalin, Mao, and the Korean War* (1993); and DAE-SOOK SUH, *Kim Il Sung: The North Korean Leader* (1988, reissued 1995).

Video documentaries. Documentary films outside of Russia, China, and North Korea almost always focus on the U.S. military experience in Korea. The following documentaries, though hardly "balanced," at least provide interpretations that compete with the American view: *The Korean War* (1992), produced by the Korean Broadcasting System (Hanguk Pangsong Kongsa); *Korea: The Unknown War* (1988), produced by Phillip Whitehead for Thames Television; and *Korea: War at the 38th Parallel* (1989), produced by the BBC (British Broadcasting Corporation). (A.R.Mi.)

Kyōto

The capital of Japan for more than 1,000 years (from 794 to 1868), Kyōto (literally, "Capital City") has been called a variety of names through the centuries—Heian-kyō ("Capital of Peace and Tranquillity"), Miyako ("The Capital"), and Saikyō ("Western Capital"), its name after the Meiji Restoration (1868) when the Imperial Household moved to Tokyo. The contemporary phrase *sekai no Kyōto* ("the world's Kyōto") reflects the interest other countries have in Japanese culture and Kyōto's own attempt to keep up with the times. Kyōto is the centre of traditional Japanese culture and of Buddhism, as well as of fine textiles and other distinctively Japanese products. The deep feeling of the Japanese people for their culture and heritage is represented in their special relationship with Kyōto—all Japanese try to go there at least once in their lives, with almost a third of the country's population visiting the city annually.

Kyōto is located on the island of Honshu some 30 miles (50 kilometres) northeast of the industrial city of Ōsaka and about the same distance from Nara, another ancient centre of Japanese culture. Gently sloping downward from north to south, the city averages 180 feet (55 metres) above sea level, and it covers an area of about 236 square miles (611 square kilometres). It is the capital of Kyōto-fu (Kyōto urban prefecture), which is at the centre of the Kinki-chihō (Kinki region). The city is also one of the centres (with nearby Ōsaka and Kōbe) of the Keihanshin Industrial Zone, the second largest urban and industrial agglomeration in Japan.

This article is divided into the following sections:

Physical and human geography

THE LANDSCAPE

The city site. On a site chosen for a new capital by the emperor Kammu, Kyōto was laid out in 794 on the model of Ch'ang-an (modern Sian), the capital of China's T'ang dynasty. The plan called for a rectangular enclosure with a grid street pattern, 3.2 miles north to south and 2.8 miles east to west. The Imperial Palace, surrounded by government buildings, was in the city's north-central section. Following Chinese precedent, care was taken when the site was selected to protect the northern corners, from which, it was believed, evil spirits could gain access. Thus, Mount Hiei (Hiei-zan; 2,782 feet) to the northeast and Mount Atago (Atago-yama; 3,031 feet) to the northwest were considered natural guardians. Mount Hiei especially came to figure prominently; between the 11th and 16th centuries warrior-monks from its Tendai Buddhist monastery complex frequently raided the city and influenced politics. The Kamo and Katsura rivers—before joining the Yodo River (Yodo-gawa) to the south—were, respectively, the original eastern and western boundaries. But the attraction of the eastern hills kept the city from filling out to its original western border until after World War II. Kyōto is actually cradled in a saucer of hills on three sides that opens to the southwest toward Ōsaka.

Climate. Kyōto is most beautiful in spring and fall. The rainy season (June–July) lasts three to four weeks, and summers are hot and humid. Winter brings two or three light snows and a penetrating "chilling from below" (*sokobie*). The yearly mean temperature of Kyōto is about 59° F (15° C); the highest monthly mean, 80° F (27° C), is in August, and the lowest, 38° F (3° C), is in January. The average yearly rainfall is about 62 inches (1,574 millimetres).

Street pattern

The city layout. The original grid pattern of the streets has been retained. Numbered avenues run east and west, Shijō-dōri ("Fourth Street") being the busiest. Karasuma-dōri, running north from the JR Kyōto station, divides the city roughly into halves. Under it is one of the two lines of the municipal subway. The other, newer line, completed in 1997, runs from the JR Nijō station in the west across the city to the east and then to Daigo southeast of the city. Kyōto was the first city in Japan to have electric streetcars (starting in 1895), which eventually made it necessary to widen the major thoroughfares to allow for city-wide service.

The historic area of Kyōto has few large factories or businesses, a fact reflected in the look of the inner city—shops and workshops, residences, and offices all standing side by side. Stringent building codes limit the height of buildings in order to preserve the overall look of the historic city. Characteristic of the architecture are tiled roofs and wood weathered to dark brown, but telephone poles (now made of concrete) and a forest of television antennas protrude at every turn. A typical Kyōto house presents a narrow and low front to the street, but as it recedes it gains in height and embellishment—all this a reflection of its past history and character: wariness of the marauding monk, the zealous revenue collector, or the curious neighbour. Rarely does one enter a home beyond the front vestibule; if one is invited in, it is good form to demur.

Because of earthquakes and conflagrations, the attacks of monks from Mount Hiei, and the Ōnin War (1467–77), which utterly destroyed the city, little of Kyōto's historical architecture predates the 17th century. Replacements and renovations, of course, followed previous plans, but the single shining example of Heian-period architecture remaining is the soaring Phoenix Hall (Hōō-dō) of the Byōdō Temple (Byōdō-in), located a few miles southeast of the city on the banks of the Uji River (Uji-gawa).

Temples and shrines

Buddhist temples and Shintō shrines abound. Their grounds and those of the Kyōto Imperial Palace (Kyōto Gosho) and Nijō Castle (Nijō-jo) give Kyōto more green areas than most Japanese cities. Kyōto claims some 1,660 Buddhist temples, more than 400 Shintō shrines, and even some 90 Christian churches. Major Buddhist institutions include East Hongan Temple (Higashi Hongan-ji) and West Hongan Temple (Nishi Hongan-ji), the former with the world's largest wooden roof of its kind and the latter containing some of the best examples of architectural and artistic expression of the Azuchi-Momoyama period (1574–1600); Ryōan Temple (Ryōan-ji), with its famous rock-and-sand Zen garden; Tenryū Temple (Tenryū-ji), in the Arashiyama district to the west; Kiyomizu Temple (Kiyomizu-dera), built on stilts on the side of the eastern hills; and Kinkaku Temple (Kinkaku-ji), the "Golden Pavilion" (burned down by a deranged student in 1950 but rebuilt exactly), and Ginkaku Temple (Ginkaku-ji), the

The main house of Katsura Detached Palace, Kyōto.
Camera Tokyo

"Silver Pavilion," both of which were products of the Ashikaga shoguns' attraction to Zen. The great Shintō shrines are Kitano, Yasaka, and Heian, the last built in 1894 to commemorate the 1,100th anniversary of Kyōto's founding.

The buildings of the Kyōto Imperial Palace, originally located farther west, date from 1855 and are re-creations, in the same "monumental" Japanese style, of earlier structures that were destroyed by fire. Nijō, built by the Tokugawa shogunate, is a "token" castle, but it contains many cultural treasures; it is known for its "chirping floors" (to signal the approach of an intruder) and elaborate wall paintings of the Kanō school. The two foremost examples of traditional Japanese landscape architecture are the Katsura Detached Palace (Katsura Rikyū) in the southwest corner of the city and the Shūgakuin Detached Palace set in the northeast hills. Katsura underwent a complete renovation using perfectly matched modern materials; its buildings are models of Japanese architectural aesthetic expression. Shūgakuin contains three gardens, the third with an artificial lake. From there one can view the entire expanse of the city stretching out to the south.

THE PEOPLE

Kyōto is one of the largest cities in Japan. Its population—which includes a sizable foreign community comprising mainly Koreans (many brought there forcibly during World War II), Chinese, and Americans—has remained relatively stable for a number of years. Most of the city's residents live in the central districts, but increasingly people are moving to outlying and suburban areas.

A major item remaining on the municipal agenda has been how to assimilate the thousands of *burakumin,* the historical outcaste group, who live in segregated communities in the city. This has been a continuing social problem largely in the older urban areas of western Japan, particularly Kyōto, Ōsaka, and Kōbe. Despite the fact that the last discriminatory legal bars were removed in 1969, social and occupational progress has lagged.

THE ECONOMY

Industry. Kyōto is a city of thousands of medium and small industries, many of them family owned and operated. Traditional handicrafts abound, and their manufacture for the tourist trade is an important element of Kyōto's economic life. The central part of the city is crowded with small workshops, which produce such typical Japanese goods as fans, dolls, Buddhist altar fittings, and lacquer ware. Antipollution measures have forced the once-thriving Kiyomizu pottery kilns to move to nearby Yamashina.

Silk weaving

For centuries silk weaving, centred in the north-central Nishijin district, has been one of Kyōto's major industries. Along with the geisha and entertainment sector, the fine textiles, delicate fabrics, and embroidery represent a continuity of Kyōto's traditional role as the centre of Japanese culture. In addition, the Fushimi district in southern Kyōto, favoured with excellent water, produces some of Japan's finest sake. Also located in southern Kyōto are several industries established after World War II that produce industrial ceramics, women's garments, and medical instruments. Since the early 1980s, companies such as Kyocera Corporation (originally called Kyōto Ceramics Co., Ltd.) have put Kyōto in the forefront of such high-technology industries as electronics, robotics, and computers. Since the early 1990s the city, like the rest of Japan, has been struggling with economic recession.

Leisure trade

Commerce. Kyōto is mainly a consumer city. It is the national centre of silk and fine textile wholesaling, but its main commercial activity is retail trade. The Gion and Pontocho districts, famed for their geisha and *maiko* (apprentice geisha), offer a variety of traditional and foreign food and drink. During the summer *yuka* (platforms on stilts) are set up on the banks of the Kamo River in the heart of town, and strolling troubadours pass below as a reminder of how kabuki theatre originated. Traditional Japanese inns (*ryōkan*) abound, and many Western-style hotels cater to the wedding, tourist, and convention trades. A large conference centre near the foot of Mount Hiei hosts major industrial exhibitions and international conferences.

Transportation. Most of Japan's east-west traffic must come through Kyōto. During the Tokugawa period (1603–1867) the city was the western terminus of the Tōkaidō, the road that connected Kyōto to Edo (now Tokyo). River traffic to Ōsaka favoured the Yodo. Today the numerous, high-speed "bullet" trains of the Shinkansen give reliable service to major cities east and west. Interurban lines between Kyōto and Ōsaka-Kōbe and Nara provide fast and frequent local service. Kyōto itself finally abandoned streetcars in the 1970s. The Meishin Expressway links Kyōto to Ōsaka and Nagoya.

ADMINISTRATION AND SOCIAL CONDITIONS

Government. Kyōto urban prefecture, which extends to the Sea of Japan, is under the administration of an elected governor, while the city is administered by an elected mayor and city council.

Education. Kyōto was traditionally organized into extended neighbourhoods, called *machi,* and after the Meiji Restoration these were designated as the administrative units for general public education; in this way, the city preceded the national effort to systematize primary education.

Higher education

Kyōto is surpassed only by Tokyo in its number of institutions of higher learning, but it claims several more Nobel Prize laureates than Tokyo, a point it reminds the latter of from time to time. The city's relatively calm atmosphere, its distance from the hurly-burly of national government, and its numerous cultural and religious institutions and facilities are cited as prime reasons for its educational advantages. There are more than 40 two-year and four-year colleges and universities with a total annual enrollment of more than 100,000 students. The state-run Kyōto University, established in 1897, is the second most prestigious school in the country. Dōshisha University, the leading private educational institution, was founded in 1875 by Niijima Jō (also called Joseph Hardy Neesima), who was the first Japanese to graduate from a Western college (Amherst College in 1870). Major Buddhist universities include Ryūkoku, Ōtani, and the smaller Hanazano.

CULTURAL LIFE

During the millennium that Kyōto served as the nation's capital and residence of the imperial family, it became the preserver of the Japanese "spirit." This is exemplified in its varied and unique cultural institutions: the schools of tea ceremony (*cha-no-yu*) and flower arranging (*ikebana*); the theatrical arts of nō, kabuki, and traditional dance; or the masterpieces of calligraphy, painting, sculpture, and architecture that can be found everywhere in the city. Kyōto is the repository of hundreds of designated "national treasures" and "important cultural objects," representing a significant proportion of the national total. Included among these are individuals who have been named "living national treasures" (*ningen kokuhō*) in recognition of their superior skills in the traditional arts and crafts.

Most of the important works of art are housed in Kyōto's temples and shrines, many of which are themselves listed as national treasures. Even institutions that do not normally display their collections periodically have public showings at which their treasures can be viewed. Kyōto also has numerous museums, including Kyōto National Museum (founded 1889), containing national treasures; Kyōto Municipal Art Museum (1933); and Kyōto Municipal Traditional Crafts Centre (1976).

The birthplace of traditional Japanese drama, Kyōto maintains an active theatrical life. Several nō stages offer frequent performances, and the annual opening performance (*kaomise*) at the Minami Theatre is the customary inauguration of the national kabuki season. A traditional form of humorous pantomime, *Mibu kyōgen*, is performed faithfully by troupes of amateurs.

The three major festivals (*matsuri*)—Aoi in May, Gion in July, and Jidai in October—are almost national events. The Jidai-matsuri ("Festival of the Ages") is a parade depicting, in period costumes, Japan's entire history. The Gion-matsuri dates from the 9th century and features more than 30 elaborate, carefully preserved, hand-drawn floats, some decorated with French Gobelin tapestries imported through Nagasaki during Tokugawa times. The northern hills—Mount Hiei with its scenic drive and the Takao district for its fall foliage—are famed for their well-tended stands of Japanese cedar (*sugi*).

History

The ancient city

Kyōto as the national capital dates from 794, although the area was settled earlier by Korean immigrants who brought with them the skills of sericulture and silk weaving. As noted above, the planned city was between the Katsura and Kamo rivers, but it soon extended beyond the eastern banks of the Kamo. The powerful Fujiwara family dominated the Heian period. Excessive Buddhist influence at the old capital of Nara had occasioned the removal of the government to Nagaoka and then to Kyōto, where the building of Buddhist temples was proscribed. As an exception, Rashōmon, the great southern gateway, was flanked by Tō Temple on the east and Sai Temple on the west; Sai Temple was short-lived, but the handsome, five-tiered pagoda of Tō Temple is a classic landmark.

Following the decline of the Fujiwara and the ascendance of the Minamoto in the late 12th century, political and military leadership was vested in a shogun ("generalissimo"), the first of whom, Minamoto Yoritomo, chose to administer the expanding domains from Kamakura to the east. It was during the Kamakura period (1192–1333) that many of the Buddhist temples were established, and indigenous sects of Buddhism, together with Zen from the continent, appeared. During the ensuing Muromachi period (1338–1573), the Ashikaga shogunate moved the government back to Kyōto. The aristocratic culture of the Heian era blended with the culture of Zen that had developed under the samurai (warriors), resulting in the refinement of the nō theatre, the tea ceremony and flower arranging, and pottery making.

By the mid-16th century, however, the city had been so devastated that St. Francis Xavier, on a pilgrimage to Kyōto, could not even locate the imperial court, much less seek an imperial audience. The city's fortunes revived under the regimes of the national unifiers Oda Nobunaga and Toyotomi Hideyoshi. Buddhists, especially the Tendai monks on Mount Hiei, were so anathema to Nobunaga that he set fire to the entire monastery complex; but under Hideyoshi, an ardent patron of the arts, Kyōto flourished. One of his tea parties was attended by thousands of people and went on for days.

With the ascendance of the Tokugawa shogunate at the beginning of the 17th century, the political centre again moved, this time to Edo (modern Tokyo). The imperial court was left to pursue its ceremonial functions, and access to it was carefully monitored. Only after the arrival of Matthew Perry in 1853 and the collapse of the shogunate did Kyōto again come to the fore. At the Nijō Temple in 1867 the last Tokugawa shogun relinquished to the imperial court his mandate to rule the nation, marking the first time in more than 200 years that a ruling Tokugawa had set foot in Kyōto.

The modern city

Shortly after the proclamation of the Meiji Restoration, however, the young Meiji emperor took up residence in the new capital, Tokyo—a move that has not been forgotten in Kyōto. Kyōto busied itself in outbidding Ōsaka to become in 1872 the site of an annual exhibition that was held for more than 30 years. During World War II U.S. Secretary of War Henry L. Stimson, recalling his visits to Kyōto, struck the city from the list of targets for aerial bombing. Its cultural treasures intact, it maintains a special place in the hearts of the Japanese and, increasingly, in the eyes of the world. At the same time, it is emerging as one of the country's centres of high-technology industry.

BIBLIOGRAPHY. On the history of Kyōto, see JOHN WHITNEY HALL, *Japan from Prehistory to Modern Times* (1970, reprinted 1991); EDWIN O. REISCHAUER and JOHN K. FAIRBANK, *East Asia: The Great Tradition* (1960); JOHN K. FAIRBANK, EDWIN O. REISCHAUER, and ALBERT M. CRAIG, *East Asia: The Modern Transformation* (1965); and EDWIN BAYRD (NED BAYRD), *Kyoto* (1974). For particular topics in the history of Kyōto, see JOHN LOWE, *Old Kyōto: A Short Social History* (2000); SUZANNE MARIE GAY, *The Moneylenders of Late Medieval Kyōto* (2001); MARY ELIZABETH BERRY, *The Culture of Civil War in Kyōto* (1994); and MAKIKO NAKANO, *Makiko's Diary: A Merchant's Wife in 1910 Kyōto* (1995). Descriptive works include GOUVERNEUR MOSHER, *Kyōto: A Contemplative Guide* (1964, reprinted 1978); DONALD KEENE, *Landscapes and Portraits: Appreciations of Japanese Culture* (1971, reissued as *Appreciations of Japanese Culture*, 1981); YOSHIKAZU IZUMOJI, *Kyōto*, 34th ed. (1983; originally published in Japanese, 1963); TADASHI ISHIKAWA, *Palaces of Kyōto* (1968; originally published in Japanese, 1962), and *Imperial Villas of Kyōto* (1970); ALEXANDRE GEORGES and DONALD RICHIE, *The Temples of Kyōto* (1995); FRANÇOIS BERTHIER and GRAHAM PARKES, *Reading Zen in the Rocks: The Japanese Dry Landscape Garden* (2000); and THOMAS WRIGHT and KATSUHIKO MIZUNO, *Zen Gardens: Kyōto's Nature Enclosed* (1996). OTIS CARY, *Mr. Stimson's "Pet City": The Sparing of Kyōto, 1945* (1975), details Kyōto's survival during World War II. On social life and customs, see RUTH L. GAINES, *City-Royal: A Memory of Kyōto* (1953). JUDITH CLANCY, *Exploring Kyōto: On Foot through the Ancient Capital* (1997); and MASON FLORENCE, *Kyoto* (1999), are good guidebooks.

(O.C./Ed.)

Lakes

A lake is a body of slowly moving or standing water that occupies an inland basin of appreciable size. Definitions that precisely distinguish lakes, ponds, swamps, and even rivers and other bodies of nonoceanic water are not well established. It may be said, however, that rivers and streams are relatively fast moving; marshes and swamps contain relatively large quantities of grasses, trees, or shrubs; and ponds are relatively small in comparison to lakes. Geologically defined, lakes are temporary bodies of water.

This article treats lake basins and sedimentation; the physical and chemical properties of lake waters; lake currents, waves, and tides; and the hydrologic balance of lakes. For information on related systems, see the article RIVERS. The place of lakes within the hydrologic cycle is further dealt with in HYDROSPHERE, as are certain aspects of lake sedimentation and water chemistry. See BIOSPHERE, THE, for information on lacustrine life-forms.

For coverage of other related topics in the *Macropædia* and *Micropædia,* see the *Propædia,* sections 222, 231, 232, 354, and 10/33, and the *Index.*

This article is divided into the following sections:

General considerations

OCCURRENCE

Within the global hydrologic cycle, freshwater lakes play a very small quantitative role, constituting only about 0.009 percent of all free water, which amounts to less than 0.4 percent of all continental fresh water. Saline lakes and inland seas contain another 0.0075 percent of all free water. Freshwater lakes, however, contain well over 98 percent of the important surface waters available for use. Apart from that contained in saline bodies, most other continental waters are tied up in glaciers and ice sheets, and the remainder is in groundwater.

Four-fifths of the 125,000 cubic kilometres (30,000 cubic miles) of lake waters occur in a small number of lakes, perhaps 40 in all. Among the largest are Lake Baikal, in Central Asia, containing about 23,000 cubic kilometres (5,500 cubic miles) of water; Lake Tanganyika (19,000 cubic kilometres [4,600 cubic miles]), in eastern Africa; and Lake Superior (12,000 cubic kilometres [2,900 cubic miles]), one of the Great Lakes of North America. The Great Lakes contain a total of about 25,000 cubic kilometres (6,000 cubic miles) of water and, together with other North American lakes larger than 10 cubic kilometres (2 cubic miles), constitute about one-fourth of the world's lake waters. The Caspian Sea, though not considered a lake by some hydrologists, is the world's largest inland sea. Located in Central Asia, the Caspian Sea has an area of about 386,000 square kilometres (149,000 square miles).

Distribution

Although lakes are to be found throughout the world, the continents of North America, Africa, and Asia contain about 70 percent of the total lake water, the other continents being less generously endowed. A fourth of the total volume of lake water is spread throughout the world in uncounted numbers of small lakes. Anyone who has flown over much of the Canadian plains area cannot help but be struck by the seemingly endless skein of lakes and ponds covering the landscape below. Though the total volume of water involved is comparatively small, the surface area of lake water is substantial. The total surface area of all Canadian lakes has been estimated to exceed the total surface area of the province of Alberta. The U.S. state of Alaska has more than 3,000,000 lakes with surface areas greater than 8 hectares (20 acres).

The larger, deeper lakes are a significant factor within the cycle of water—from rain to surface water, ice, soil moisture, or groundwater and thence to water vapour. These lakes receive the drainage from vast tracts of land, store it, pass it on seaward, or lose it to the atmosphere by evaporation.

This so-called water budget of lakes is important enough to have warranted considerable study throughout the world, with each lake or lake system possessing its own hydrologic idiosyncrasies. Aside from being of scientific interest, water-budget studies serve to reveal the dependence of each lake on particular hydrologic factors, thus enabling better management practices. These may include restrictions on water utilization during drought conditions, dike construction and evacuations prior to flooding, control of water levels to ensure efficient power production, and major decisions associated with diversions of watercourses in order to enhance water quantity and quality.

On a local basis, even the smaller lakes play an important hydrologic role. The relatively high ratio of exposed surface area to the total water volume of these lakes accentuates their effectiveness as evaporators. In some cases the efficiency of lakes in losing water to the atmosphere is locally undesirable, because of public and industrial requirements for lake water. A striking example of this condition is the Aral Sea, located in Central Asia. Although it is still one of the world's largest bodies of inland water, in the second half of the 20th century its area was reduced by two-fifths and its mean surface level had dropped by more than 12 metres (40 feet), as a result primarily of the diversion of the Syr Darya and Amu Darya rivers for irrigating adjoining fields. In some basins (*e.g.,* the Chad basin in Africa), lakes are the terrestrial end point of the hydrologic cycle. With no outflow downstream toward the oceans, these closed lakes swell or recede according to the balance of local hydrologic conditions.

USES AND ABUSES OF LAKES

Importance to man

In today's industrial societies, requirements for water—much of which is derived from lakes—include its use for dilution and removal of municipal and industrial wastes, for cooling purposes, for irrigation, for power generation, and for local recreation and aesthetic displays. Obviously, these requirements vary considerably among regions, climates, and countries.

In another vein, it is convenient to use water to dilute liquid and some solid wastes to concentrations that are not intolerable to the elements of society that must be exposed to the effluent or wish to use it. The degree of dilution that may be acceptable varies from situation to situation and is often in dispute. In some cases, dilution is used purely to facilitate transport of the wastes to purification facilities. The water may then be made available for reuse.

Lake water is also used extensively for cooling purposes. Although this water may not be affected chemically, its change in thermal quality may be detrimental to the environment into which it is disposed, either directly, by affecting fish health or functions, or indirectly, by causing excessive plant production and ultimate deoxygenation due to biological decay. Both fossil- and nuclear-fuelled power plants are major users of cooling water. Steel mills and various chemical plants also require large quantities.

Concern with thermal pollution of surface waters is concentrated principally on rivers and small lakes. With power requirements in modern societies increasing by about 7 percent per year, however, some apprehension has been expressed about the future thermal loading of even the largest lakes. It was predicted that thermal inputs to each of the North American Great Lakes will increase by nearly 11 times during the last three decades of the 20th century. In terms of energy to be disposed in this fashion, the numbers are staggeringly large. These lakes have such large volumes, however, and such large surface areas (from which much of the heat goes into the atmosphere) that there is some question about the nature and magnitude of the actual effects.

The economic importance of waterways as communication links is enormous. In the earliest times, when travel by many societies was substantially by water, travel routes became established that resulted in relationships between cultural factors and surface hydrology networks. Today, river and lake systems serve as communication links and play an important role in shipping because of the large cargo capacities of merchant vessels and the still fairly uncongested condition of inland waterways. Oceanic shipping lanes play the major role, but river and lake systems, which link inland ports with the oceans, have been key factors in the rates of economic growth of many large inland ports.

Commercial fisheries and other food industries reap great harvests from the major lakes of the world. The quality of the fish catch has steadily decreased, however, as a result of pollution in many lakes, with the more desirable species becoming less plentiful and the less desirable species gradually dominating the total. Other commercial harvests from lakes include waterfowl, fur-bearing mammals, and some plant material, such as rice.

Each of the uses described has associated with it the means for abuse of the very characteristics of lakes that make them desirable. Wise management of natural resources has never been humankind's forte. Municipalities and industries have polluted lakes chemically and thermally, the shipping that plies large inland water bodies leaves oil and other refuse in its wake, water used for irrigation often contains chemical residues from fertilizers and biocides when it is returned to lakes, and the populace that so desperately demands clean bodies of water for its recreation often ignores basic sanitary and antipollution practices, to the ultimate detriment of the waters enjoyed.

Major problems

Among the major problems affecting the optimum utilization and conservation of lake waters are eutrophication (aging processes), chemical and biological poisoning, and decreases in water volumes. In the former case, discussed in more detail later, the enrichment of lakes with various nutrients supports biological productivity to an extent in which the ultimate death and decay of biological material places an excessive demand on the oxygen content, resulting in oxygen depletion in the worst cases. Phosphates and nitrates are two of the types of nutrients that are most important in this connection, particularly since they are often introduced in critical quantities in waste effluents from human sources. Other examples of chemical pollution of lakes include the introduction of DDT and other pesticides and heavy metals such as mercury. Bacteriological contamination of lake waters resulting in levels that constitute a hazard to health is another common result of disregard for the environment.

Water-quantity problems are complex, being related to natural vagaries of supply and levels of consumptive utilization of water. In the latter case, the percentage of water returned to the source after utilization varies with the use. The largest losses are due to actual water diversions and processes that result in evaporative losses. The use of large quantities of lake water for cooling purposes by industry and utilities, for example, may raise lake temperatures near the effluents sufficiently to cause increased evaporation. The use of certain types of cooling towers results in even larger losses. Some of the water evaporated will stay within the lake basin, but some will be lost from it.

Another example of this type of loss is connected with the possible application of weather-modification techniques to alleviate the heavy lake-effect snowfalls experienced along the lee shores of large lakes in intermediate latitudes. Redistribution of precipitation always raises the possibility of redistribution of water among various basins.

Lake-effect snowfall is just one example of the influence of lakes on local climate. The ability of large bodies of water to store heat during heating periods and to lose it more gradually than the adjacent landmasses during cooling periods results in a modifying influence on the climate. Because of this propensity, a lake cools air passing over it in summer and warms air passing over it in winter. Consequently, the predominantly downwind side of a lake is more influenced by the ameliorating effects of a lake.

In most instances, moisture is also passed to the atmosphere. In summer, lake cooling serves to stabilize the air mass, but winter heating tends to decrease stability. The moisture-laden, unstable winter flows off lakes produce so-called snowbelts, which affect downwind cities. The snowbelts are usually of limited extent, often within about a kilometre of the lake shore. (R.K.L./Ed.)

Lake basins

CLASSIFICATION OF BASINS

Limnology

The name given to the study of lakes is limnology. Limnologists have used several criteria for the development of systems for classifying lakes and lake basins but have resorted particularly to the mechanisms that have produced lake basins. These have been summarized and examined in *A Treatise on Limnology*, by the American limnologist G.E. Hutchinson, which includes treatment of tectonism, volcanism, landslides, glaciation, solution, river action, wind action, coastline building, organic accumulation, animal activity, meteoritic impact, and human activity.

Basins formed by tectonism, volcanism, and landslides. Tectonism—or movements of the Earth's crust—have been responsible for the formation of very large basins. During the late Miocene (the Miocene Epoch includes the time interval from 26,000,000 to 7,000,000 years ago), broad, gentle earth movements resulted in the isolation of a vast inland sea across southern Asia and southeastern Europe. Through most of the Tertiary Period (from 65,000,000 to 2,500,000 years ago), sub-basins developed that gradually were characterized by a great range of salinities. Resumption of communication with the oceans occurred later, and there is evidence of considerable variation in water levels. The present remnants of these inland bodies of water include the Caspian Sea and the Aral Sea, along with numerous smaller lakes. The Black Sea, which was also once part of this large inland basin, is now in direct communication with the oceans.

In some cases, elevated land areas may already contain depressions that eventually form lake basins. Lake Okeechobee, Florida, is cited as being such a basin, formed by uplift of the ocean floor.

Tectonic uplift may interfere with natural land-drainage patterns in such a way as to produce lake basins. The Great Basin of South Australia, some of the lakes in Central

Africa (*e.g.*, Lakes Kioga and Kwania), and to some extent Lake Champlain, in the northeastern United States, are examples of this mechanism. Land subsidence, due to earthquake activity, also has resulted in the development of depressions in which lakes have evolved. Many such cases have been reported within the past 300 years.

The damming of valleys as a result of various tectonic phenomena has resulted in the formation of a few lake basins, but faulting, in its great variety of forms, has been responsible for the formation of many important lake basins. Abert Lake, in Oregon, lies in the depression formed by a tilted fault block against the higher block. Indeed, many lakes in the western United States are located in depressions formed through faulting, including Lake Tahoe, in the Sierra Nevada, California. Great Salt Lake, Utah, and other nearby salt lakes are remnants of a large Pleistocene lake, Lake Bonneville, which was formed at least partly by faulting activity (the Pleistocene Epoch includes the interval from about 2,500,000 to about 10,000 years ago).

In other parts of the world, faulting has also played an important role in basin formation. Lake Baikal and Lake Tanganyika, the two deepest lakes in the world, occupy basins formed by complexes of grabens (downdropped faulted blocks). These lakes are among the oldest of modern lakes, as are other graben lakes, particularly those within the East African rift system, which extends through the East African lake system and includes the Red Sea (see further CONTINENTAL LANDFORMS: *Rift valleys*).

Basins formed from volcanic activity are also greatly varied in type. The emanation of volcanic material from beneath the surface can be explosive, or it can issue in a gentle and regular manner. This range of activity and the variation of types of material which may be involved produce many different types of basins.

Crater lakes

One broad category includes those occupying the actual volcanic craters or their remnants. Crater lakes may occupy completely unmodified cinder cones, but these are rare. Craters caused by explosions or by the collapse of the roofs of underground magma (molten silica) chambers and those caused by explosion of new volcanic sources and that are built of nonvolcanic material are other examples. The latter are termed maars, following the local name for such forms in Germany. They are found, however, in several locations, including Iceland, Italy, and New Zealand. The maars of the volcanic district of Eifel, Germany, are among the best-known of these formations.

The collapse of magma chambers and the development of very large surface craters called calderas is an important source of lake basins. Crater Lake, Oregon, is a typical example, exhibiting characteristically great depth and a high encircling rim. Some caldera basins evolved with gently sloping sides, however, due to the deposition of material from a series of explosions and a gentler collapse of the structure. Secondary cones may develop within calderas, as shown by Wizard Island, in Crater Lake. The largest caldera in the world, which contains Lake Toba, in Sumatra, was formed through a combination of volcanic action and tectonic activity. Lake Toba's basin is contained in a strike-slip fault belt along the entire length of the Barisan Mountains of Sumatra. A vast, initial eruption of lava under gas pressure collapsed the magma reservoir, forming a depression that filled with water, producing the lake. Renewed volcanic activity subsequently led to the formation of an island in the centre, but a second collapse later cut it in two. Additional tectonic activity has further modified the lake's configuration.

Lake basins may also arise from the action of lava flows that emanate from volcanic fissures or craters. Lake Mývatyn, in Iceland, was formed in a basin arising from the collapse of the interior part of a large lava flow. Other basins have formed as the result of volcanic damming. This usually happens where a lava flow interrupts the existing drainage pattern.

Lava damming

Lake basins also may form following the blockage of a drainage depression by landslides. These may be temporary in nature because of the eroding action of the lake on the damming material. Lake Sārez in the Pamirs is stable, being dammed by a rockslide.

Basins formed by glaciation. The basin-forming mechanism responsible for the most abundant production of lakes, particularly in the Northern Hemisphere, is glaciation. The Pleistocene glaciers, which seem to have affected every continent, were especially effective in North America, Europe, and Asia. The retreat of ice sheets produced basins through mechanical action and through the damming effect of their ice masses at their boundaries.

In some cases, lakes actually exist in basins made of ice. In other cases, water masses may form within ice masses. Such occurrences are rare and are not very stable. Damming by ice masses is a more common phenomenon but is also likely to be relatively temporary. Glacial moraine (heterogeneous sedimentary deposits at glacier margins) is also responsible for the occurrence of dammed lake basins. The Finger Lakes of New York state are dammed by an endmoraine.

Ice-scour lakes, cirque lakes, and fjords

Ice sheets moving over relatively level surfaces have produced large numbers of small lake basins through scouring in many areas. This type of glacial rock basin contains what are known as ice-scour lakes and is represented in North America by basins in parts of the high Sierras and in west central Canada (*e.g.*, near Great Slave Lake). Tens of thousands of these lakes are found in the ice-scoured regions of the world. Many of them are interconnected with short streams and may contain narrow inlets. Characteristically, they may be dotted with numerous islands and sprawling bays. Many are comparatively shallow. Where they are particularly abundant, they may cover up to 75 percent of the total surface, as in Quetico canoe country of Minnesota.

Glacier scouring associated with the freezing and thawing of névé (granular snow adjacent to glacier ice) at the head of a glaciated valley may produce a deepened circular basin termed a cirque. These are found in widely scattered mountain locations. The action of glaciers in valleys can produce a similar type of basin, often occurring in series and resembling a valley staircase. Ice movement from valleys through narrow openings has produced another type of rock basin, known as glint lake basins, particularly in Scandinavian regions.

Piedmont and fjord (*i.e.*, a river valley that has been "drowned" by a rise of sea level) lakes are found in basins formed by glacial action in long mountain valleys. Excellent examples are found in Norway, the English Lake District, the Alps, and the Andes. In North America, several regions contain this type of lake basin. In British Columbia, many good examples exist, the largest of which are the Okanagan and Kootenay systems. These are long, narrow lakes of substantial depth. In northwestern Canada, some of the largest lakes, including Lake Athabaska, Great Slave Lake, and Great Bear Lake, are of this type, although they are not found in the same type of mountainous terrain. These lakes, as well as the North American Great Lakes, resulted from the movements of large ice sheets that deepened existing valleys.

The Wisconsin (latest stage of Pleistocene glaciation) ice sheet was responsible for shaping the present Great Lakes system, which drains mainly eastward to the Atlantic through the St. Lawrence River, during its retreat. The principal stages in the history of these lakes have received much study, and several stages of retreat and advance of the ice sheet have been identified. Behind the lobes of the ice sheet, ice lakes developed that drained according to the modifications of preexisting valleys for glacial action. As the mass of ice retreated far to the north, glacial rebound (uplift of the Earth's crust in response to removal of the loading by ice) caused a general tilting of the land surface; the new lake basins also contributed to the subsequent changes through their own erosional action.

The material comprising glacial moraines or glacial outwash may provide dams that confine postglacial waters. The Finger Lakes, in New York state, constitute one interesting group of this type. These lakes were formed through glacial scouring of existing valleys, which were blocked at both the northern and southern ends by morainic deposits.

A variety of basin types have been formed in the different types of glacial drift deposits, including basins in morainic material, kettle lakes, channels formed by water movement in tunnels beneath the ice masses, and lake basins formed by thawing in permafrost. An interesting example of glacial action is the formation of giant's kettles; these are glacial potholes in the form of deep cylindrical holes. Their origin is still uncertain. Sand, gravel, or boulders are sometimes found at their bottom. The kettles vary from a few centimetres to a metre or more in diameter. Good examples are found in the Alps, Germany, Norway, and in the United States.

Formation of natural dams

Basins formed by fluvial and marine processes. Fluvial action in several forms can produce lake basins; the most important processes include waterfall action, damming by sediment deposition from a tributary (fluviatile dams), sediment deposition in river deltas, damming by tidal transport of sediments upstream, changes in the configuration of river channels (*e.g.,* oxbow lakes and levee lakes), and solution of subsurface rocks by groundwater.

This last mechanism has produced the well-known formations in the Karst region, in Yugoslavia, which include subterranean and surface cavities and basins in limestone. The term karstic phenomena is applied to similar cases in many parts of the world (see further CONTINENTAL LANDFORMS: *Caves and karst landscape features*). Solution lakes in Florida (*e.g.,* Deep Lake) are also of this origin, as are Lünersee and Seewlisee, in the Alps. Other rock types susceptible to solution basin formation include gypsum and halite. Mansfeldersee in Saxony was formed in this manner.

In some coastal areas, longshore marine currents may deposit sufficient sediment to block river outflows. This damming action may be of varying intensity, and it may also occur in lake regions, where such current action causes sediment deposition that leads to the formation of multiple lakes. Accumulation of organic plant material can also result in structures that produce lake basins; Silver Lake, Nova Scotia, evolved from damming by plant material. Structural formations of coral are another potential cause of damming.

Basins formed by wind action, animal activity, and meteorites. Wind action may lead to dam or dune construction or erosion and thus can play a role in lake-basin formation. The latter case has been demonstrated in North America; a number of basins in Texas and northward, on the plains east from the Rocky Mountains, are thought to have originated from wind erosion—at least in part. Moses Lake in Washington state was formed by windblown sand that dammed the basin.

Mammals have constructed lake-forming dams; the American beaver is highly skilled at this, and its activities in this connection have established it as a symbol of industriousness. Man has also been busy in this regard and is fully capable of producing lakes that would rival the largest of the more natural variety. Plans once proposed for the damming of the Yukon River in Alaska would, if carried through, result in the formation of a lake larger than Lake Erie in surface area. Other human activities, such as quarrying and mining, also have produced cavities suitable for lake formation.

The last major mechanism of basin formation is that due to meteoritic impact. Meteorite craters are best preserved in arid climates and are often dry for this reason. A few lakes are known in craters, however, including Ungava Lake, in Quebec. In many other cases, it has not been possible to definitely confirm that basins that have the appearance of meteorite craters have, indeed, been produced by meteorite impact. Controversial ones include the bay lakes of southeast North America.

TOPOGRAPHY OF BASINS

Lakes meet with both the atmosphere and the underlying material of their terrestrial basins and interact with each. The topography and configuration of the lake bottom and the nature of the bottom materials vary considerably. They are of sufficient importance to most lake processes to warrant recognition as basic lake characteristics.

The surface area of a lake can easily be determined by cartographic techniques, but lake-volume determinations require knowledge of lake depths. Throughout the world, lakes important enough to warrant study have been sounded, and many nations have completed comprehensive programs to determine the bathymetry of large numbers of lakes. Lake sounding involves traversing a lake to collect either point or continuous measurements of depth until an accurate survey is made. Modern sounding devices measure the time taken for emitted sound to return after reflection from the bottom, relying on a knowledge of the speed of sound in water. The more sophisticated of these also provide for detection of the depths of stratification in sedimentary materials on the lake bottom. The employment of laser devices from aircraft is a recent development that is based on the transmission of light beams with wavelengths that will penetrate water.

For more practical purposes, lake morphology is a stable characteristic. Shore erosion, sediment deposition and transfer, and other processes, however, including dredging by man, may significantly alter a lake's bottom topography and thus affect navigation, currents, and ecological factors, such as fish spawning grounds.

SEDIMENTS AND SEDIMENTATION

Lake sediments are comprised mainly of clastic material (sediment of clay, silt, and sand sizes), organic debris, chemical precipitates, or combinations of these. The relative abundance of each depends upon the nature of the local drainage basin, the climate, and the relative age of a lake. The sediments of a lake in a glaciated basin, for example, will first receive coarse clastics, then finer clastics, chemical precipitates, and then increasingly large amounts of biological material, including peats and sedges.

Clues to lake history

Geologists can deduce much about a lake's history and the history of the lake basin and climate from the sedimentary records on its bottom. A sediment core contains such clues as ripple marks caused by current or wave action, carbonaceous layers, and alternations of strata that include cold- and warm-water species of fossils, pollen, and traces of chemicals of human derivation. These data provide the basis for extensive documentation of lake history (paleolimnology). Some well-known historical events, such as major volcanic eruptions, the clearing of North American forests by early settlers as revealed by pollen concentrations, the first extensive use of certain heavy metals by industry, and nuclear explosions, provide reference points in the sediment record.

Many of the materials that are detrimental to the ecology of a lake—*e.g.,* excessive quantities of nutrients, heavy metals, pesticides, oil, and certain bacteria—are deposited in lake sediments by chemical precipitation or the settling of particulate matter. These materials are potentially available for regeneration into the lake water and must be considered in any planning for measures to abate lake pollution. Within the uppermost lake sediments, large volumes of interstitial water are often present. This water may have high concentrations of nutrients and other constituents and enhance the exchange potential with the lake proper.

Clastic sediments. Waters draining into a lake carry with them much of the suspended sediment that is transported by rivers and streams from the local drainage basin. Current and wave action along the shoreline is responsible for additional erosion and sediment deposition, and some material may be introduced as a result of wind action. Rivers and streams transport material of many different sizes, the largest being rolled along the riverbed (the bed load). When river water enters a lake, its speed diminishes rapidly, bed-load transport ceases, and the suspended load begins to settle to the bottom, the largest sizes first. Lake outlets carry with them only those materials that are too small to have settled out from the inflows or those that have been introduced adjacent to the outflow. Because dynamic processes that keep materials suspended are generally more active near the shore, lake sediments are usually sorted by size; the rocks, pebbles, and coarse sands occur near shore, whereas the finer sands, silts, and muds are, in most cases, found offshore.

Clastic material over most of a lake basin consists principally of silts and clays, especially away from shores and river mouths, where larger material is deposited. Clays exist in a variety of colours, black clays containing large concentrations of organic matter or sulfides and whiter clays usually containing high calcium carbonate concentrations. Other colours, including reds and greens, are known to reflect particular chemical and biological influences.

Organic sediments are derived from plant and animal matter: *förna* is recognizable plant and animal remains, *äfja* finely divided remains in colloidal suspension, and *gyttja* is a deposit formed from *äfja* that has been oxidized. Rapid accumulation of organic matter in still lakes is not uncommon; in the English Lake District, 5 metres (15 feet) of lake sediment of organic origin accumulated over a period of about 8,000 years. Pollen analysis has been used to accurately decipher climatic conditions of the lake in the past.

Varved deposits are the product of an annual cycle of sedimentation; seasonal changes are responsible for the information. Varves are a common feature in many areas and especially so where the land has received meltwaters from ice sheets and glaciers. The deposits consist of alternating layers of fine and coarse sediments.

Coarse clastic materials seldom are larger than boulders (25 centimetres [10 inches]), and the type of material in sizes larger than silt and clay frequently reveals its source. Materials along lakeshores can in most cases be traced back to a particular eroded source within the local drainage basin, and the distribution of this material provides evidence of the predominant current or wave patterns in the lake.

Volcanic ash is deposited downwind from its source. Ash from volcanic activity during the Pleistocene Epoch can often be dated and used as a stratigraphic marker. Lakes throughout the northwestern United States contain some of the best examples (the Mazama ash), and one deposit in the central United States, called the Pearlette ash deposit, occurs in beds as thick as 3 metres (10 feet).

Deposition of salts

Chemical precipitates. The major chemical precipitates in lake systems are calcium, sodium, and magnesium carbonates and dolomite, gypsum, halite, and sulfate salts. Calcium carbonate is deposited as either calcite or aragonite when a lake becomes saturated with calcium and bicarbonate ions. Photosynthesis can also generate precipitation of calcium carbonate, when plant material takes up carbon dioxide and bicarbonate and raises the pH above about 9 (the pH is a measure of the acidity or alkalinity of water; acid waters have a pH of less than 7, and the pH of alkaline waters range from 7 to 14).

Dolomite deposition occurs in very alkaline lakes when calcium carbonate and magnesium carbonate combine. Recent dolomites have been found in Lake Balkhash, in Kazakhstan. In many saline lakes, gypsum deposition has occurred; Lake Eyre, Australia, is estimated to contain more than 4,000,000,000 tons of gypsum. For gypsum to be deposited, sulfate, calcium, and hydrogen sulfide must be present in particular concentrations. Hydrogen sulfide occurs in deoxygenated portions of lakes, usually following the depletion of oxygen resulting from decomposition of biological material. Bottom-dwelling organisms are usually absent.

Lakes that contain high concentrations of sodium sulfate are called bitter lakes, and those containing sodium carbonate are called alkali lakes. Soda Lake, California, is estimated to contain nearly 1,000,000 tons of anhydrous sulfate. Magnesium salts of these types are also quite common and can be found in the same sediments as the sodium salts. Other salts of importance occurring in lake sediments include borates, nitrates, and potash. Small quantities of borax are found in various lakes throughout the world. Lakes with high alkalinity levels, such as Mono Lake in California, can still support some forms of life.

The gradual increase of sediment thickness through time may threaten the very existence of a lake. When a lake becomes shallow enough to support the growth of bottom-attached plants, these may accelerate the extinction of a lake. In several European countries, steps are being taken to restore lakes threatened by choking plant growth. Lake Hornborgasjön, Sweden, long prized as a national wildlife refuge, became the subject of an investigation in 1967. Lake Trummen, also in Sweden, was treated by dredging its upper sediments. In Switzerland, Lake Wiler (Wilersee) was treated by the removal of water just above the sediments during stagnation periods.

Lake waters

CHEMICAL COMPOSITION

Although the chemical composition of lakes varies considerably throughout the world, owing to the varying chemistry of the erosion products of different lake basins, in most cases the principal constituents are quite similar. Human influences also have contributed substantially to the chemical composition of lakes, and, although industrial effluents vary somewhat from lake to lake, many of the chemical effects of human activities are similar throughout the world. Another source in the chemical balance of lakes is the dissolved and suspended material contained in precipitation. Again, human activities have been in large part responsible for steadily increasing concentrations of this input.

Concentration of salts

Salinity, nutrients, and oxygen. Salinity is the total concentration of the ions present in lake water and is usually computed from the sodium, potassium, magnesium, calcium, carbonate, silicate, and halide concentrations. Several important bodies of inland waters, often called inland seas, have very high salinities. Great Salt Lake, in Utah, has a salinity of about 200,000 milligrams per litre, as compared with Lake Superior's value of about 75 and an estimated mean for all rivers of 100 to 150. These ions are steadily introduced to lakes from rivers and rainwater, where they concentrate because of the evaporative loss of relatively pure water.

Where inflowing rivers erode igneous rocks, lake salinity values are relatively low, but, where soluble salts are available for erosion, salinities are relatively high. In general, it has been found that, of the cations (positively charged ions), calcium concentrations are highest, followed by magnesium, sodium, and potassium, in that order. Among the major anions (negatively charged ions), carbonate is generally the most abundant, followed by sulfate and chloride.

Other inorganic ions, though present in smaller concentrations, are of great importance. In particular, the nutrients (especially phosphate, nitrate, and silicate), heavy metals (*e.g.,* mercury, manganese, copper, lead), and polychlorinated hydrocarbons (DDT, for example) have attracted recent interest because of their role in ecological problems. Although sources of nutrients and mercury exist that are not directly related to human activities, budget studies and studies of the historical records available in sediment cores clearly reveal the great impact of human disposal of these constituents in lakes. Rainfall and dry fallout are small but significant chemical inputs to lakes. The release of gases and particulate matter into the atmosphere from factories and similar sources has increased dramatically in recent years, with consequent alterations in the chemistry of rainwater. It has been estimated, for example, that 16,000 tons of nitrogen, about 8 percent of the total from all sources, is introduced annually to Lake Erie from atmospheric action.

Oxygen concentration

The substance of most interest in lakes is oxygen; once introduced to the lake water, its concentration is subject to factors within the water. Biological production (photosynthesis) releases oxygen into the water, while biological decay consumes it. Various chemical reactions within the lake system also affect the concentration of dissolved oxygen. The main source is the passage of oxygen through the air-water interface, which is affected principally by the lake temperatures; at low temperatures the partial pressure of dissolved oxygen in water is reduced. Consequently, during cold seasons, especially when vertical mixing is greatly enhanced because of a lack of thermal structure and increased wind stirring, lakes are replenished with oxygen. In the warmer seasons, although surface waters may remain more or less saturated and even supersaturated, the concentrations are lower. Beneath the surface, oxygen consumption through biological decay may cause serious depletion. Oxygen depletion also occurs near the bottom because of processes at the mud-water interface, many of which are still inadequately explained.

In winter months, a rapid formation of ice or the establishment of strong winter thermal stratification may significantly inhibit the replenishment of oxygen. Where ice cover lasts for long periods, a loss of oxygen at the mud-water interface may have repercussions for the whole lake, particularly if density currents cause significant vertical transport.

In tropical regions, where the winter replenishment mechanism (turnover) is absent, there is great reliance on the occasional occurrence of cold spells or on significant nighttime cooling to promote oxygen replenishment. Deep lakes in these regions are often anoxic (lacking in oxygen) in the deeper portions.

At any particular time, lake waters or waters entering a lake may have a biological or chemical potential for oxygen utilization. Measurements of this are termed BOD (biological oxygen demand) or COD (chemical oxygen demand). These concepts are used as partial indicators of the quality of waters being introduced to a lake.

Lakes that have a vertical salinity gradient strong enough to prevent winter turnover will usually be deoxygenated at depths where the vertical diffusion of oxygen is less than the oxygen demand. Such lakes are termed meromictic.

Carbon dioxide. Another gaseous substance of great importance that is exchanged with the atmosphere at the surface is carbon dioxide. Photosynthesis requires the presence of carbon dioxide, and it is released during biological breakdown.

Relation of carbon dioxide to pH and carbonate solubility

Carbon dioxide is very soluble in lake water; it forms carbonic acid, which dissociates and raises the concentration of hydrogen ions (lowering the pH). The relative proportions of bicarbonate, carbonate, and free carbon dioxide depend upon the pH. At high values of pH, carbonate ions will predominate; at low values, free carbon dioxide and carbonic acid will predominate.

Various carbonates (particularly sodium, calcium, potassium, and magnesium) are important to the carbon dioxide system. Increased pressure of carbon dioxide in the system increases the solubilities of these carbonates. In some cases, photosynthetic activity results in precipitation of certain carbonates. The entire carbon dioxide system and its behaviour at various pH values is very complex but can be interpreted from historical knowledge of lake sediments.

In waters that are neither very acidic (pH much less than 7) nor very basic (pH much greater than 7 but less than 14), the carbon dioxide system serves as a buffer, because, within limits, a change in pH will cause a shift within the system that ultimately serves to offset the pH change. Consequently, most lakes have a pH between 6 and 8. Some volcanic lakes are extremely acidic, however, with pH values below 4, and some lakes with very high pH values, such as Lake Nakuru, Kenya, also occur in nature.

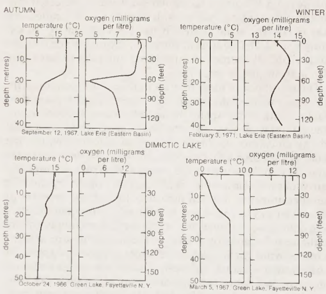


Figure 1: Seasonal variation of temperature and dissolved oxygen in dimictic (having two circulation periods annually) and meromictic (undergoing incomplete circulation at the fall overturn) lakes (see text).

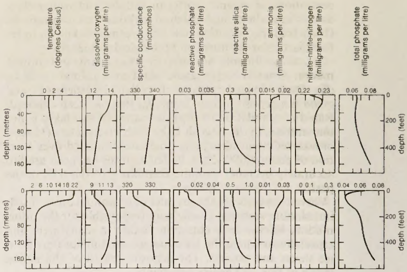


Figure 2: Distribution of some physical and chemical variables in central Lake Ontario in (bottom) July 1969 and (top) June 1970.

Sulfates, nitrates, and phosphates. Sulfate usually occurs as a principal ion in lake waters. Under anaerobic conditions in which bacteria persist in the oxidation of biological material, hydrogen sulfide is produced. When anoxic conditions exist in the deep waters just above the sediments, and the water is acidic enough to precipitate the iron present, hydrogen sulfide occurs. The characteristic and unpleasant odour of this gas is often popularly identified with the "death" of a lake. Big Soda Lake, Nevada, is extremely rich in this substance.

Nitrogen and its various compounds form another complex system in lakes, appearing as free nitrogen in solution, organic compounds, ammonia, nitrite, and nitrate. Sources of nitrogen compounds include influents to the lake (the most important source), fixation in the lake, and precipitation. Losses are experienced mainly through effluents but also by denitrification, sediment formation, and loss to the atmosphere.

Orthophosphate and various organic phosphates are the most important phosphorous compounds in lakes. Phosphates and nitrates are heavily consumed in the upper portion of lakes during periods of high productivity of phytoplankton. Increased concentrations occur in deeper portions due to decay of falling biological material and regeneration from the sediments, especially during anoxic conditions or stormy periods in shallow lakes. As limiting nutrients in many lake productivity cycles, phosphates and nitrates are often identified as controllable elements in situations where abatement is necessary to control eutrophication. Carbon is also a necessary constituent for production and in some cases can be the limiting component. Because carbon is less easily controlled and not often limiting, however, phosphates are most frequently named as substances to be reduced in effluents from industry and municipalities.

Silica also is present in lake waters, and, as with the other nutrients, it is introduced in influents and to some extent from the sediments. The production of diatom blooms is a major process for reducing silicate concentrations. Within this context, silica can also be regarded as a limiting nutrient.

THERMAL PROPERTIES

Basic physical data

Pure water freezes at 0° C (32° F), boils at 100° C (212° F), and has a latent heat of evaporation of 539.55 calories per gram, a latent heat of sublimation (ice) of 679 calories per gram, and a specific heat of 1.01 calories per gram, per ° C, at 0° C. The temperature of maximum density of water at atmospheric pressure occurs at 3.94° C (39.09° F). At the freezing point, ice has a lower density than water. For natural waters with high salinities, such as the oceans and inland seas, each of the values above is significantly altered. In most lakes, however, these numbers are quite representative.

The density of water increases at pressures above one atmosphere (the pressure at sea level). Thus, pure water at

10°C (50° F) has a density of 0.9995 at one atmosphere and 1.0037 at the pressure existing at a lake depth of 1,000 metres (3,000 feet). Water raised from great depths to conditions of lower pressure experiences adiabatic cooling (without significant heat exchange with surrounding water), but there are very few lakes in which this factor can be of much significance.

Vertical mixing and overturn. It is useful to know how the temperature of maximum density changes with depth (*e.g.*, from 3.94° C at the surface to 3.39° C at 500 metres depth [38.10° F at 1,500 feet]). The fact that the temperature of maximum density of most lake waters is close to 4° C (39° F), whereas ice forms at temperatures close to 0° C in response to surface cooling, vertical mixing takes place. When density increases with depth, the lake is said to be stable. Unstable conditions exist when density decreases with depth. Cooling at the surface to temperatures below 4° C establishes stability based on a negative thermocline (a positive thermocline is a vertical decrease in temperature with depth), because density will increase with depth. Ice then forms at the surface, enabling liquid water to exist beneath the ice in lakes, unless they are shallow enough to freeze to the bottom.

During the warming season, after ice has melted, heating increases the density of the surface waters, causing them to sink until stability is achieved. When surface heating proceeds above the temperature of maximum density, this process ceases, and the vertical thermal structure maintains and strengthens its stable condition, based on a positive thermocline. Turnovers tend to be seasonal.

Dimictic, monomictic, holomictic, and meromictic lakes

Mixing due to cooling or warming processes that increase the density of surface waters sufficiently to cause them to sink results in what is termed circulation, or overturn, of lake waters. Lakes that cool to below 4° C in winter experience two turnover periods, as just described, and are called dimictic lakes. Most lakes in temperate regions fall into this category. Lakes that do not cool to below 4° C undergo overturn only once per year and are called warm monomictic. Lakes that do not warm to above 4° C also experience only one overturn period per year and are called cold monomictic. There are many examples of the former, including lakes in the tropical regions and generally as far north as about 40°. The cold monomictic type, however, is less common but can be found at high latitudes and at high altitudes (in the Alps, for example).

All the types described that circulate at least once throughout are called holomictic. It is possible, however, for lakes to be stable despite the thermal processes that normally induce overturn owing to the existence of a positive salinity gradient with depth (chemocline). This type is called meromictic, and, in those cases where stability is permanent in at least part of the lake, the deep waters do not experience overturn and consequently are deoxygenated. Three principal origins of meromixis have been recognized. Ectogenic meromixis results from either the intrusion of seawater into a lake, as in the case of flooding from an unusually high sea level (*e.g.*, Hemmelsdorfersee, in Germany), or the introduction of fresh water through land drainage and precipitation to a saline lake (*e.g.*, Soda Lake, Nevada). Crenogenic meromixis is due to the introduction of saline water by springs, and biogenic meromixis is due to the uptake of salts from the lake sediments. North American examples include Lake Mary, Wisconsin, and Sodon Lake, Michigan.

A strong vertical salinity gradient that exists in the upper portion of a lake will affect the thermal structure by inhibiting the downward mixing of heat. In holomictic lakes, however, the downward mixing of heat due primarily to wind action usually compresses or concentrates the thermocline until it essentially separates an upper layer (epilimnion) from a lower layer (hypolimnion), each possessing weak or nonexistent vertical thermal gradients. The thermocline normally begins to grow at the beginning of the warming season. As summer passes and autumn commences, it intensifies and deepens. The onset of the cooling sees the beginning of the decay of the thermocline from above, although it usually continues to deepen until it is completely destroyed. The process just described is commonly found in lakes in temperate regions and is a seasonal phenomenon. During any period of strong warming, one or more shallower thermoclines may be observed to develop and move downward to the seasonal thermocline.

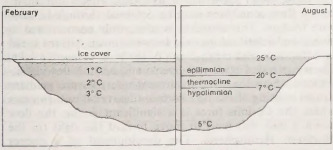


Figure 3: Temperature distribution in lakes in the middle latitudes of North America in winter and summer.

The heat budget of lakes. The heat budget of a lake includes several major factors: net incoming solar radiation, net exchange of long-wave radiation emitted by the lake surface and the atmosphere, transfer of sensible heat at the surface interface, and latent-heat processes. Those processes that are usually of much smaller importance include net inflow and outflow of heat advected by streamflow, precipitation, and groundwater flow, conduction from terrestrial heat flow, and dissipation of kinetic energy. In some cases, however, river inflow may be of more importance, such as where flow is from a nearby glacier or where the volume inflow is a significant fraction of the lake volume. Within a large lake the heat-budget considerations for a particular location must also take into account the local advection of heat within the lake by currents.

Seasonal variations in incoming solar radiation

Incoming solar energy varies seasonally and with the latitude and is greatly influenced by cloud cover. The fraction that is reflected away from the lake surface depends upon the solar angle, the turbidity of the atmosphere, and the wave state, or surface roughness. In middle latitudes this ranges from about 6 percent in summer months to about 14 percent in winter.

The amount of radiation emitted by the lake surface is proportional to the fourth power of the surface temperature, whereas the radiation emitted by the clouds and atmosphere overlying the lake depends primarily upon the amount and height of the clouds and the temperature and moisture content of the atmosphere near the lake surface.

The fluxes of sensible heat and moisture at the lake surface are of great importance yet are still poorly understood. They depend upon the vertical gradients of temperature and vapour pressure above the water, respectively, and upon the factors that influence the transfer processes, such as wind and atmospheric stability. The transfer of sensible heat may be either into or out of the lake surface, usually on a seasonal basis but also sometimes on a diurnal basis. It is also possible but less likely for condensation to occur on a lake surface.

Heat flow through the bottom of lakes is normally of small significance, but exceptions exist. In a very deep lake where low rates of heating are important, such as Lake Baikal, Russia, the results may be detectable. In some ice-covered lakes where other sources of heating are small, heat flux through sediments also has been shown to be significant.

The dissipation of wind energy that has been transferred to water movements is quite insignificant, as is the effect of heat transfer due to chemical and photosynthetic processes.

In latitudes and altitudes where ice is a factor, the latent heats of fusion and of evaporation of ice must also be considered within any heat-budget considerations. Heat-balance studies have been performed for lakes that are always ice-covered. Solar radiation is often an important factor where ice thickness and consistency permits penetration. The heat balance of the ice is often difficult to assess, as long-wave radiation and evaporation factors are not easily measured and are very important. The exchange of sensible heat may not be large during summer months in these cases but is likely to be significant in the colder months. Several lakes that are ice-covered have been shown to be meromictic; two examples are Lake Tuborg, Ellesmere Island and Lake Bonney, Antarctica.

Heat-balance measurements or estimates have been made for many lakes throughout the world. Results show that the difference between the highest and lowest heat content for each lake varies from around 5,000 calories per square centimetre for high and low latitudes to around 45,000 calories for some midlatitude lakes.

Heat balance of two representative lakes

The relative importance of each of the major terms of the heat budget is shown by data for two North American lakes: Lake Ontario, a large, deep, middle-latitude lake; and Lake Hefner, a relatively small, shallow lake in Oklahoma. The energy unit frequently employed is the langley (one gram calorie per square centimetre), and the figures given are approximate monthly means of langleys per day. Net solar radiation input to Lake Ontario varies from 80 to 600 (Lake Hefner varies from 200 to 600), midwinter to midsummer. Net losses due to long-wave radiation from Lake Ontario are nearly 100 throughout the year (Lake Hefner varies from 100 to 200). Evaporation losses for Lake Ontario vary from 250 in midwinter to slightly negative values in early summer (Lake Hefner varies from 450 in late summer to 150 in spring). Conduction of heat from the surface of Lake Ontario varies from 250 in winter to about minus 100 in summer (Lake Hefner varies considerably from 80 to −80 for the same time interval).

Heat added to a lake at the surface is usually mixed mechanically downward as a result of wind action. This process keeps the upper portion of a lake relatively uniform thermally. Consequently, a thermal gradient (thermocline) becomes established between the upper mixed layer (epilimnion) and the deep portion of the lake (hypolimnion). In shallow lakes or shallow portions of large lakes, the thermocline will eventually intercept the lake bottom so that no hypolimnion exists. Normally, as the heating season progresses, the thermocline intensifies and deepens. Secondary thermoclines may develop in the epilimnion, and these will migrate downward to the main seasonal thermocline. On very warm, still days, a thin surface layer may store heat before a mixing episode transfers heat downward. When the cooling season commences, the mixing that tends to destroy the thermocline is enhanced by vertical convection. If the cooling continues until the entire thermocline is eliminated, the lake becomes essentially isothermal and no longer exhibits the characteristics of a two-layered system.

When a lake is stratified, the most important process for downward transfer of heat to the hypolimnion is through eddy conduction. The coefficient of eddy conductivity is determined empirically and varies substantially from lake to lake. Mixing processes are generally more active in coastal areas, so that isotherms can be expected to slope downward toward shore. In large, relatively unprotected lakes, wind stress at the surface causes convergence or divergence or both of shallow waters along coastlines. Isotherms will slant upward toward the shore, and hypolimnion water may even become exposed at the surface. These occurrences are of great importance with regard to the distribution of heat within stratified lakes.

Thermal pollution in lakes

Heat introduced to lakes in large quantities, as a waste product of cooling processes in power-generating plants and other industrial concerns, is presently viewed with some concern as a pollutant, especially in small lakes. If the heat is injected at the surface it will spread initially according to the momentum of the influent and the speed and direction of ambient surface currents. When the initial momentum is sufficiently dissipated, the heat will spread mainly as a consequence of turbulent mixing processes. Throughout these events, substantial losses of heat to the atmosphere may occur, so that the full effects of the thermal input are not borne solely by the lake. Temperature values at the surface, adjacent to the influent-heat source, may be raised to a very high level—as much as several centigrade degrees. Under certain conditions fish-activity tolerances may be exceeded, and undesirable algae and plankton production may be stimulated.

If waste heat is not released at the surface but is diffused over a large depth range or injected at depth, the large local-surface-temperature problem is avoided. Losses to the atmosphere in this case, however, are also greatly reduced, and the net heat input to the lake as a whole is much greater. Over a long period, this may prove to be more detrimental to the general ecology than near-surface injection.

Lake hydraulics

CURRENTS

The principal forces acting to initiate water movements in lakes are those due to hydraulic gradients, wind stress, and factors that cause horizontal or vertical density gradients. Lake water movement is usually classified as being turbulent.

Hydraulic effects are frequently the result of inflows and outflows of water. These may be substantial and continuous or weak and sporadic; in terms of the ratio of the volume of the inflow or outflow to the lake volume, the latter is the most frequently observed situation.

The stress of wind moving over the lake surface causes a transport of water within the lake, as well as the movement of energy downwind through the mechanism of surface waves. The wind is therefore one of the most important external forces on a lake. It can be relatively consistent in speed and direction, or it can be highly variable in either or both.

Pressure gradients. Water movements can occur as a result of internal pressure gradients and from density gradients caused by variations in temperature, sediment concentration, or the concentration of dissolved substances. Surface water in lakes can become denser than underlying water either by cooling or heating, because the temperature of maximum density for pure lake water is about 4° C (39° F). Water entering a lake from rivers with a high concentration of dissolved substances will sink to a lake level of similar density. These movements are both horizontal and vertical, but the net effect is downward, if not vertical, motion.

Horizontal pressure gradients can result from many different processes that act to produce density gradients. One example is the situation of solar heating in a shallow nearshore region, where the heat is committed to the warming of a relatively small volume of water. This produces a water of lower density than the near-surface water of an adjacent deep region, where the heat is spread throughout a greater volume. Consequently, the pressure gradient force will act to move the warmer water offshore and to replace it from below with cooler water.

Influence of the Coriolis effect

Lake currents are the result of complex interactions of forces, but in many cases a small number of particular forces dominate. In the case of horizontal flow in the absence of horizontal pressure gradients, assuming no friction, water set in motion will curve to the right in the Northern Hemisphere because the Earth rotates from west to east. This effect is called the Coriolis force, and it will continue to influence water motion until there is a balance with the centrifugal force. This movement causes free-floating markers to move in an elliptical manner with a period that depends upon the latitude. In Lake Ontario, for example, it is about 17 hours. Where a dominating pressure gradient exists, the balance of the pressure-gradient force with the Coriolis force results in the so-called geostrophic flow, at right angles to the pressure gradient, with low pressure on the left (Northern Hemisphere). These conditions are most nearly realized only in very large lakes and in the oceans.

In those small lakes where hydraulic effects dominate, steady flow conditions may be achieved through balance with friction. This situation is commonly encountered in rivers, and relationships exist between mean current speed and the slope and mean depth of the river or narrow lake. These are called gradient currents and occur following situations where the wind or atmosphere pressure gradient causes a tilting of the lake surface (denivellation). In cases where the Coriolis force is a significant factor, the flow down a lake will tend to move toward the right (in the Northern Hemisphere). The development of a deeper countercurrent to the left will occur to compensate for the piling up of water on the right side.

Horizontal pressure gradients will be important in lakes where there are significant inflows of water with markedly different density from ambient lake density or where significant differential surface heating occurs.

Wind stress. Currents resulting from wind stress are the most common in lakes. Considerable research is still underway into the mechanism of transfer of wind momentum to water momentum. The stress on the lake is proportional to some power of wind speed, usually taken to be 2, although it evidently varies with wind speed, wave conditions, and atmospheric stability. In large, deep lakes, away from the boundaries, where wind-stress effects may be balanced by Coriolis-force effects, theory suggests that the surface current will move in a direction 45° to the right of the wind and that deeper currents are progressively weaker and directed farther to the right. The depth at which flow is opposite to the wind direction is effectively the depth below which there is no influence from the wind. This depth, designated *D*, can theoretically occur at about 100 metres (300 feet) in large, deep, midlatitude lakes. Observations show varying degrees of fidelity to theory because of complications from coastal effects and thermal stratification.

In coastal regions, if water depth is a significant fraction of or greater than *D*, winds blowing parallel to the shore will transport water either onshore or offshore. In the latter case, where the coast is to the left of the wind flow (Northern Hemisphere), the water driven offshore is replaced by cooler, deeper water (upwelling).

Effect of shearing forces

Internal waves and Langmuir circulation. Under stratified conditions a strong thermocline will essentially separate a lake into two layers. Shearing forces that develop between these layers cause a motion, termed internal waves, that may serve to directly dissipate a substantial portion of a lake's kinetic energy and act as a coupling between motion in the epilimnion and hypolimnion. A great range of periodicities is observed in the oscillations of the thermoclines, particularly in large lakes. Internal seiches, which are responsible for relatively long-period internal waves, are discussed later.

A small-scale circulation phenomenon that has aroused considerable attention on lakes is Langmuir circulation. On windy days, parallel "streaks" can be observed to develop on the water surface and exhibit continuity for some distance. These streaks may be caused by convergence zones where surface froth and debris collect. Langmuir circulation thus appears to be a relatively organized mixing mechanism wherein sinking occurs at the streaks and upwelling occurs between the streaks. Under favourable circumstances, this appears to be a key process for mixing heat downward in lakes.

SURFACE WAVES

Wind blowing over a calm lake surface first produces an effect that may appear as a widely varying and fluctuating ruffling of the surface. The first wave motion to develop is relatively regular, consisting of small, uniformly developed waves called capillary waves. These are quite transient, dissipating rapidly if the wind dies away or developing to the more commonly observed and more persistent gravity waves.

Energy will be continually fed to the waves by the frictional drag of the air moving over the water and by the direct force of the wind on the upwind face of the waves. The latter effect occurs only while the waves move more slowly than the wind. Pressure differences at the air–water interface also contribute energy to surface waves. Energy losses occur due mainly to turbulence in the water and, to a smaller extent, to the effects of viscosity.

Waves will continue to grow as long as there is a net addition of energy to them. Their height will increase as a function of wind speed and duration and the distance over which it blows (fetch). Most lakes are so small that fetch considerations are unimportant. Studies in larger lakes, however, have shown that the height of the highest waves are related to the fetch. In these lakes, waves as high as several metres are common, although waves of about seven metres (23 feet) are the highest to be expected. Wave heights in a given portion of a lake may vary considerably, due to interactions that suppress some waves and amplify others. As waves develop, their lengths increase, even after their height has stopped increasing. The phenomenon of swell, commonly observed in the oceans, is not truly realized, even in the largest lakes.

Growth and movement of waves

Waves travel in the same direction as the wind that generated them and at right angles to their crests. If they meet a solid object rather than a sloping beach, much of their energy will be reflected. If they enter shallow water obliquely, they are refracted. Wave speed, for waves longer than four times the depth of the water, is approximately equal to the square root of the product of the depth and the gravitational acceleration. For waves in relatively deep water, the wave speed is proportional to the square root of the wavelength.

As wave height increases, the sharpening of the wave crest may result in instability and a breaking off of the crest, a process hastened by the wind. This results in the familiar whitecaps. Waves that run ashore break up in surf. The wave height first decreases slightly, then increases, and the speed decreases, and eventually the wave form disappears as it crumbles into breakers. These can be plunging forms, in which the top curls right over the forward face, or of the spilling type, in which the crest spills down the forward face. A particular wave may break several times before reaching shore. (R.K.L.)

BIBLIOGRAPHY. MARY J. BURGIS and PAT MORRIS, *The Natural History of Lakes* (1987); and LAURENCE PRINGLE *et al.*, *Rivers and Lakes* (1985), are two treatments written for the general reader. G. EVELYN HUTCHINSON, *A Treatise on Limnology*, vol. 1, *Geography, Physics, and Chemistry* (1957, reissued 1975), still a classic work, contains a comprehensive treatment of the physical and chemical aspects of lakes and includes an excellent bibliography of previous work. Less fundamental but more recent summaries of scientific lake studies include GERALD A. COLE, *Textbook of Limnology*, 4th ed. (1994); CHARLES R. GOLDMAN and ALEXANDER J. HORNE, *Limnology* (1983); ROBERT G. WETZEL, *Limnology*, 2nd ed. (1983); ABRAHAM LERMAN (ed.), *Lakes—Chemistry, Geology, Physics* (1978); and WERNER STUMM (ed.), *Chemical Processes in Lakes* (1985). Up-to-date books on hydrology and hydrogeology are R. ALLAN FREEZE and JOHN A. CHERRY, *Groundwater* (1979); PATRICK A. DOMENICO and FRANKLIN W. SCHWARTZ, *Physical and Chemical Hydrogeology* (1990); and RAFAEL L. BRAS, *Hydrology* (1990). The sedimentology of lakes is the subject of J. MCMANUS and R.W. DUCK (eds.), *Geomorphology and Sedimentology of Lakes and Reservoirs* (1993). Prediction of the behaviour of lakes is dealt with in B. HENDERSON-SELLERS, *Engineering Limnology* (1984); and A. JAMES (ed.), *An Introduction to Water Quality Modelling*, 2nd ed. (1993). The dynamics of lakes are covered by K. HUTTER (ed.), *Hydrodynamics of Lakes* (1984); JAMES LIGHTHILL, *Waves in Fluids* (1978); and JEROME WILLIAMS and SAMUEL A. ELDER, *Fluid Physics for Oceanographers and Physicists* (1989). An up-to-date book on general freshwater biology is BRIAN MOSS, *Ecology of Fresh Waters*, 2nd ed. (1988). Also of interest on this subject are the studies by C.S. REYNOLDS, *The Ecology of Freshwater Phytoplankton* (1984); and by MAX M. TILZER and COLETTE SERRUYA (eds.), *Large Lakes: Ecological Structure and Function* (1990). The consequences of building large-scale dams are analyzed in EDWARD GOLDSMITH and NICHOLAS HILDYARD, *The Social and Environmental Effects of Large Dams* (1986). KENT W. THORNTON, BRUCE L. KIMMEL, and FORREST E. PAYNE (eds.), *Reservoir Limnology* (1990), analyzes the ecology of reservoirs in their similarities to and differences from lakes. The effects of pollution on lakes are discussed in B. HENDERSON-SELLERS and H.R. MARKLAND, *Decaying Lakes: The Origins and Control of Cultural Eutrophication* (1987); E.B. WELCH and T. LINDELL, *Ecological Effects of Wastewater: Applied Limnology and Pollution Effects* (1992); THEODORA E. COLBORN *et al.*, *Great Lakes, Great Legacy?* (1990); and WILLIAM ASHWORTH, *The Late, Great Lakes: An Environmental History* (1986). Continuing research can be found in *Limnology and Oceanography* (bimonthly). (R.K.L./Ed.)

Land Reform and Tenure

The concept of land reform has varied over time according to the range of functions which land itself has performed: as a factor of production, a store of value and wealth, a status symbol, or a source of social and political influence. Land value reflects its relative scarcity, which in a market economy usually depends on the ratio between the area of usable land and the size of that area's population. As the per capita land area declines, the relative value of land rises, and land becomes increasingly a source of conflict among economic and social groups within the community.

The patterns of wealth and income distribution and of social and political influence are partly determined by the laws governing land tenure. These laws specify the acceptable forms of tenure and the privileges and responsibilities that go with them. They define the land title and the extent to which the owner can freely dispose of it and of the income accruing from its use. In this sense, the form of tenure determines the wealth and income distribution based on the land: if private ownership is permitted, class differentiation is unavoidable; in contrast, public ownership eliminates such distinctions. The forms of tenure range from temporary, conditional holding to ownership in fee simple, which confers total unencumbered rights of control and disposal over the land.

This article is divided into the following sections:

The purposes of land reform

Historically, land reform meant reform of the tenure system or redistribution of the land ownership rights. In recent decades the concept has been broadened in recognition of the strategic role of land and agriculture in development. Land reform has therefore become synonymous with agrarian reform or a rapid improvement of the agrarian structure, which comprises the land tenure system, the pattern of cultivation and farm organization, the scale of farm operation, the terms of tenancy, and the institutions of rural credit, marketing, and education. It also deals with the state of technology, or with any combination of these factors, as shown by recent reform movements, regardless of the political or ideological orientation of the reformers.

OBJECTIVES OF REFORM

Reform is usually introduced by government initiative or in response to internal and external pressures, to resolve or prevent an economic, social, or political crisis. Thus reform may be considered a problem-solving mechanism. The true motives for reform, however, may well differ from those announced by the reformer. The distinction between the real and proclaimed objectives may be especially significant if the proclaimed objectives have been forced upon reformers who do not support those objectives. The reformers may proclaim certain objectives merely to appease the peasants, to undermine opposition, to win international backing, or to safeguard their own positions. The proclaimed purposes of land reform, however, will be the point of departure in this article.

Political and social objectives. The most common proclaimed objective of land reform is to abolish feudalism, which usually means overthrowing the landlord class and transferring its powers to the reforming elite or its surrogates. If "foreigners" happen to be among the landlord class, the objectives become the defeat of imperialism and the end of foreign exploitation.

Another common objective is to free the peasants from subjugation to and dependence on the exploiters and make them active citizens by restoring what assertedly had been taken away from them.

A third objective is to create democracy—a stated purpose of both capitalist and Communist reformers. Most capitalist reforms are based on the premise that individual private ownership in the form of independent family farms will promote and sustain democratic institutions.

Communist reformers, in contrast, usually aim at overthrowing both feudalism and capitalism on the premise that, as a means of production, private ownership of land inherently breeds exploitation. In practice, this means "returning land to the tillers" and creating a classless, democratic society. A more immediate and practical goal of Communist reformers has often been to rally the peasants in support of the new order and against the former regime.

Finally, reform may be introduced simply as the most expedient way to resolve a crisis or avoid a revolution. The reformer, in this case, will introduce and implement just enough reform to appease the peasants and contain the conflict. This happens especially when the reformers are still in sympathy with the landlord class and consciously prefer a moderate rather than a radical reform.

These political objectives tend to undergo change during the period of implementation and are, therefore, kept vague enough to permit flexibility and modification as conditions change.

All land reforms emphasize the need to improve the peasants' social conditions and status, to alleviate poverty, and to redistribute income and wealth in their favour. They try to create employment opportunities and education and health services and to redistribute the benefits to the community at large, the younger generation as the main target.

Economic objectives. Economic development has become a major objective of governments and political parties in recent decades. Efforts have been made to encourage agricultural progress by means of agrarian reform in favour of the peasant who does not own his land or whose share of the crop is relatively small, and who therefore has little incentive to invest capital or expend effort to improve the land and raise productivity. Another mechanism has been to encourage labour-intensive cultivation, on the assumption that traditional or feudal landowners often use their land extensively and wastefully.

An equally important economic objective is to promote optimum-scale farming operations. Excessively large farms (*latifundia*) and excessively small farms (*minifundia*) tend to be inefficient. Therefore, reform aims at creating farms of optimum size given the land quality, the crop, and the level of technology.

Finally, reform aims at coordinating agriculture with the rest of the economy. In their quest for economic development and industrialization, reformers attempt to make the rural sector more responsive to the needs of the industrial sector for labour, food, industrial raw materials, capital, and foreign currency. These functions are often expected to be performed simultaneously.

TYPES OF REFORM

Whether it is called land reform or agrarian reform, the operational concept covers five main types of reform, classified according to whether they deal with land title and terms of holding, land distribution, the scale of operation, the pattern of cultivation, or supplementary measures such as credit, marketing, or extension services.

Title settlement and terms of holding land

Reforms concerned with the title to land and the terms of holding reflect a transition from tradition-bound to formal and contractual systems of landholding. Their implementation involves property surveys, recording of titles, and provisions to free the landholder from restrictions or obligations imposed by tradition. Property surveys are conducted wherever land is held by a tribe or clan or where reallocation of cultivable land routinely follows tradition. In these situations the landholder may lack the incentive to improve the land because the right of disposal belongs to the tribe, clan, or feudal lord, as in medieval Europe and in parts of present-day Africa and the South Pacific islands. Such reform affects landholding in at least three ways: it may increase security of tenure and hence incentives; it may reorganize the system of inheritance in favour of offspring; and it may bring land onto the market so that land transactions become possible. This reform, however, has little immediate effect on the scale of operation, but it does facilitate future land concentration and fragmentation. In countries where the terms of holding and tenancy are regulated by tradition, reform may seek to convert tenancy into a contractual agreement that offers some protection to the tenant and more security and incentive to improve the land and advance technology, as in Japan, India, and Pakistan.

The most common type of reform involves the redistribution of land titles from one individual to another, from individuals to a group or community at large, or from a group to individuals. The land of one landlord may be redistributed to many individuals, as in Egypt, Iran, or Ireland. Or the land of individuals may be reallocated in favour of the community at large by abolishing private ownership, as in Cuba and China. Or, again, public land may be distributed to individuals, as in various parts of Latin America.

The impact of redistribution on the scale of operations and on marketability of the land depends on the form it takes and the restrictions attached to it. If the redistributed farm was previously operated as a unit, its division means fragmentation and reduction of scale; however, if it was operated in fragments by tenants, transfer of title to the tenants would not affect the scale. The final results depend on the measures taken to prevent adverse effects.

Efforts to improve the scale of operation

Land-tenure reform, of course, can improve the scale of operations by enlarging the farm or by reducing it. Enlargement applies when the holding is increased in size, either by adding to it or by consolidating its fragmented parts. Farm consolidation involves reallocation of the total farmland within a region by land exchange, sale, or lease such that no one loses and all gain by increasing efficiency. The scale of operations may be increased by pooling resources, as in farm cooperatives and collectives that offer facilities otherwise inaccessible to a small farm.

An equally common approach is to divide large, extensively cultivated farms into smaller and more intensively cultivable units. Reduction of the scale, however, has potential problems since it may result in excessively small units or in the breakup of efficiently run farms. Operations below the optimum level may inhibit improvements in technology, capital investment, and diversification.

Changes in the pattern of cultivation relate directly to cultivation, land yield, and labour productivity. While other types of reform may influence productivity indirectly by enhancing security of tenure and the scale of operation, improvements in the pattern of cultivation affect productivity directly, through advances in technology, improved irrigation, and the application of chemicals.

Technological advance usually implies mechanization, although it may be biological and organizational only, as in crop rotation, reconditioning of the soil, improved seeding, or better utilization of available technology. The state of technology determines the level of productivity or the ratio between outputs and inputs. More advanced technology permits the cultivation of more land per unit of labour, deeper plowing, better timing of farm operations, reclamation of areas previously inaccessible, and possibly wider diversification of the crops than previously attainable. By easing the physical burden of farm work, it helps to conserve human energy. Improved technology may also be the most direct way to modify tradition without an open confrontation or a political revolution. Mechanization and advanced technology may, of course, cause displacement of labour, unemployment, or the absorption of capital at the expense of other sectors, at least in the short run; in the long run, the positive effects tend to prevail.

Improvements in irrigation include increasing the water supply, draining swampy land, and regulating the quantity and quality of water flow. Irrigation is especially important in that it involves large investments and infringes on tenure rights, both matters that invite public responsibility and intervention. Irrigation and technology are closely related to the use of fertilizer and other chemicals. Chemicals may be difficult to apply without irrigation, and neither may be practical unless farming technology has advanced beyond relatively primitive methods. Improvement of the pattern of cultivation may be inhibited, however, by traditional attitudes, the lack of skills, or the scarcity of capital. Another difficulty is that changes in the pattern of cultivation are usually long-term investments that may be too slow to satisfy immediate pressures for reform.

Measures that supplement land reform

Many improvements and changes may have to be implemented in areas outside the immediate sphere of agriculture, such as credit, marketing, and education. Unless the farmer is able and ready to take advantage of new opportunities and his product can be marketed profitably, reform efforts may be futile. Costly or inaccessible credit and the excessive charges of middlemen increase the relative costs of farming. Therefore, supervised credit, subsidies, and low-interest loans that help to replace traditional sources of credit have been common, and credit and marketing cooperatives and market regulation have been used to protect farmers against exploitation by middlemen.

Finally, improvements in general education are essential in any reform that involves the modernization of agriculture. Extension services, literacy promotion programs, the teaching of home economics, and vocational training are of special importance in helping the young and unemployed and in providing skilled labour for industry.

EVALUATION AND CRITERIA OF SUCCESS

Agrarian reform is a complex process of directed change, and its effects touch society in many ways. Therefore its evaluation may be difficult because the various social, political, and economic objectives may be inconsistent with each other. Even champions of reform and planners may have different ideas about it. Moreover, there are no generally accepted criteria for determining the success of such a program, nor adequate tools for measuring its progress.

Economic criteria. Economic indicators and criteria are basically the requisites of economic development. Economic development may be defined as a sustained increase in and achievement of a given level of per capita real income. To be sustained, the rise of per capita real income must be accompanied by changes in the economic and social structures of society, increase in total investment (capital formation), higher productivity, and full employment.

Capital formation in agriculture implies that more resources will be put at the disposal of the farmer in the form of machinery, fertilizers, and irrigation and drainage facilities, all of which contribute to the productive capacity and productivity of the land and the worker. Because capital formation partly depends on domestic saving, a higher rate of capital formation may thus be an indicator of the success of the reform in aiding economic development.

Another indicator is change in land yield or labour productivity. A rise in yield or productivity implies higher efficiency, better use of resources, and an advance in the state of technology. It also suggests an increase in the level of income and potential saving and investment. In fact, the increase in productivity may be the most important single indicator of the contribution made by reform to economic development. Change in the level of rural employment or unemployment provides another indicator, which should be reflected in the level and distribution of income.

Finally, an important indicator may be the change in agriculture's responsiveness to the demands of industry and manufacturing. The ability of agriculture to provide labour, food, industrial raw materials, and a market for industrial products is a significant measure of its contribution to industrialization and development.

Social criteria. Social and political accomplishments are more difficult to measure. One of the important indicators may be the degree of peasant participation in activities such as voting, representation, and decision making. Social and political stability, or the tendency to change governments by constitutional and nonviolent means and continuity of the social and political order without resort to force, are other indicators.

But these various indicators can only suggest that change has taken place. The results depend on the magnitude and relevance of the change. Assuming that measurement is feasible, three approaches to evaluation may be followed: the goal achievement approach, the perceived achievement approach, and the closing-the-gap (integrative) approach.

Goal achievement considers a program successful to the extent that it realizes the goals specified prior to the reform. Probably the most common economic goal is maximization, according to which efficiency dictates that reform should continue up to the point where the marginal benefit of reform is equal to its marginal cost. Land distribution in this case will continue up to the point where its net benefit is zero. Another common goal is to realize incremental gains as the reform proceeds. The reform would be successful to the extent that the effects have been positive. A 20-percent increase in capital formation in a capital-poor economy, however, may be more significant than a 20-percent increase in a capital-rich economy. Similarly, a 20-percent decline in the number of dissatisfied peasants in a peaceful or democratic society may be more conducive to stability and harmony than a 20-percent decline in a radical or violent society. Hence, this approach requires that a critical minimum achievement be specified as a criterion of success in each situation. Or, as a third alternative, the evaluation may compare the results with targets proclaimed in advance. The degree of success will be the extent to which those targets have been realized. Problems, however, will arise, especially when the goals happen to be contradictory or change over time.

Methods of assessment

Perceived achievement considers reform successful if the relevant parties perceive their goals as having been satisfied. One of the main objectives of reform has been to reduce conflict and promote harmony, both of which depend on whether a person or group perceives its expectations as fulfilled, whether it has hope that these expectations will be fulfilled, and whether it is able to express these expectations openly. The evaluation, therefore, is primarily subjective, and the net impact can be assessed only by synthesizing the views of the parties affected, including those who might be satisfied to have their losses minimized, just as much as others would want their gains maximized.

The closing-the-gap approach considers a reform successful to the extent that it closes the gap between the sector subject to reform and the more advanced sectors in society; in other words, reform would be expected to help integrate agriculture with the rest of the economy and the rural population with the urban community in terms of opportunities and levels of living. In this sense the reform would remove the dualism between the agrarian and the nonagrarian, between the technologically backward and the technologically advanced, and thus would increase labour mobility in response to the demands of the economy and development.

HISTORY OF LAND REFORM

The ideas and principles discussed so far may be illustrated by a selective survey of the history of land reform.

Greek system and reforms

Ancient reforms. The recorded history of reform begins with the Greeks and Romans of the 6th and 2nd centuries BC, respectively. Land in ancient Athens was held in perpetuity by the tribe or clan, with individual holdings periodically reallocated according to family size and soil fertility. Population increase, expansion of trade, growth of a money economy, and the opening up of business opportunities eventually made financial transactions in land an economic necessity. Land itself continued to be inalienable, but the right to use the land could be mortgaged. Thus, peasants could secure loans by surrendering their rights to the product of the land, as "sale with the option of redemption." Lacking other employment, the debtor continued to cultivate the land as *hektēmor,* or sixth partner, delivering five-sixths of the product to the creditor and retaining the rest for himself. Mortgaged land was marked by *horoi,* or mortgage stones, which served as symbols of land enserfment. When Solon was elected archon, or chief magistrate, *c.* 594 BC, his main objective was to free the land and destroy the *horoi.* His reform law, known as the *seisachtheia,* or "shaking-off the burdens," cancelled all debts, freed the *hektēmoroi,* destroyed the *horoi,* and restored land to its constitutional holders. Solon also prohibited the mortgaging of land or of personal freedom on account of debt.

The impact of the reform was extensive but of short duration. The *hektēmoroi* were freed, but since no alternative sources of support or credit were provided and creditors were uncompensated, dissatisfaction and instability persisted. Two decades of anarchy were followed by a revolution, *c.* 561 BC, that brought Peisistratus to power. He enforced the reform and distributed lands of his adversaries (who were killed or exiled) among the small holders. He also extended loans to aid cultivation and prevent migration to the city and expanded silver mining to create employment. Although the amount of land redistributed is unknown, Peisistratus was apparently able to satisfy the peasantry, secure their loyalty, and stay in power for life, but the economic effects are too vague to evaluate.

The reforms of the Gracchi

The Roman reform by Tiberius and Gaius Gracchus came between 133 and 121 BC. The land reform law, or *lex agraria,* of Tiberius was passed by popular support against serious resistance by the nobility. It applied only to former public land, *ager publicus,* which had been usurped and concentrated in the hands of large landholders. Land concentration reduced the number of owners and hence the number of citizens and those eligible to serve in the army. In addition, such concentration was accompanied by a shift from cultivation to grazing, which reduced employment and increased the poverty of the peasants, producing a crisis. The motives of the reformers continue to be debated, but it would appear that concern for the poor and political stability were major factors.

The *lex agraria* specified minimum and maximum individual landholdings, with an allowance for male children of the family. Excess land would be expropriated and compensation paid for improvements. A standing *collegium,* or commission, was to enforce the law, but implementation was delayed because Tiberius was killed in the year of its passage. When Gaius was elected tribune about a decade later, he revived the reform and went even further. He colonized new land and abolished rent on small holdings since rent on large holdings had been suspended as compensation for expropriation. Gaius was killed in 121 BC, however, and within a decade the reform was reversed: private acquisition of public land was legalized, the land commission was dissolved, rent on public land was abolished, all holdings were declared private property, and squatting on public land was prohibited. Even colonization was ended, and colonies established by Gaius were broken up. Another period of land concentration was inaugurated.

Modern European reforms. The French Revolution brought a new era in the history of land reform. Reform meant dealing with survivals of the medieval tenures that had left a common heritage in most European countries and, through them, in the colonies. The measures and approaches varied from place to place and period to period.

The French Revolution and after

On the eve of the Revolution, French society was polarized, with the nobility and clergy on one side and the rising business class on the other. The middle class was relatively small, especially in the rural areas. The majority of the peasants were hereditary tenants, either *censiers,* who paid a fixed money rent, or *mainmortables,* or serfs, who paid rent in the form of labour services, *corvée,* of about three days a week. The peasants paid various other feudal dues and taxes, from which the nobility and clergy were exempted. The Revolution overthrew the ancien régime and the feudal order and introduced land reform.

The reform repealed feudal tenures, freed all persons from serfdom, abolished feudal courts, and cancelled all payments not based on real property, including tithes. Rents based on real property were redeemable. Once the law had been passed, however, the peasants seized the land and refused to pay any rents or redemption fees; in 1792 all payments were finally cancelled. Land of the clergy and political emigrants was confiscated and sold at auction, together with common land. The terms of sale, however, often favoured the wealthy, which may explain the rise of a new class of large landowners among the supporters of Napoleon.

The social and political objectives of the reformers were fully realized. The *censiers* and serfs became owners. Feudalism was destroyed, and the new regime won peasant support. The economic effects, however, were limited. Incentives could not be increased substantially since the peasants already had full security of tenure prior to the reform. The scale of operations was not changed; and no facilities for credit, marketing, or capital formation were created. The major achievements were the reinforcement of private, individual ownership and perpetuation of the small family farm as a basis of democracy. The small family farm has characterized French agriculture ever since.

There were other reforms in most European countries. England resolved its land problems by the enclosure movement, which drove the small peasants into the towns, consolidated landholdings, and promoted large-scale operation and private ownership. Sweden and Denmark pioneered between 1827 and 1830 by peacefully abolishing village compulsion, or imposed labour service, and the strip system of cultivation, by consolidating the land, and by dividing the commons among the peasants. Though influenced by the French Revolution, only after the 1848 revolutions did Germany, Italy, and Spain free the peasants and redistribute the land. Reform in Ireland took a whole century before substantive results were achieved, in the mid-1930s, after Ireland was divided into Northern Ireland and the Irish Free State. The tenants were converted into owners by subsidized purchase of the land.

Russian reforms

The first major Russian reform was the emancipation of the serfs in 1861. At the time of emancipation about 45 percent of the land was private property and the remainder was held as allotment land, cultivated in units averaging 9.5 acres (3.8 hectares) by the peasant serfs against rent in kind and labour, payable to feudal lords. In contrast, fewer than 1,000 noble families owned about 175,000,000 acres (70,000,000 hectares) and received rent therefrom. Conflict between such extremes of poverty and wealth caused restlessness among the peasants and rendered reform inevitable. As Tsar Alexander II put it: "It is better to abolish serfdom from above than to await the day when it will begin to abolish itself from below."

The Emancipation Act of 1861 abolished serfdom and distributed allotment land among the peasants. The homestead became hereditary property of the individual, but the field land was vested in the village *mir* as a whole. The peasant paid redemption through the village authority, while the landlord received state bonds as compensation equal to 75 to 80 percent of the land market value. Though legally freed, the private serf had to ransom his freedom by surrendering a part of the allotment land. In contrast, serfs belonging to the imperial family were emancipated in 1863 and received the maximum amount of land fixed by law. Serfs of the state were emancipated in 1866 and allowed to keep the land they occupied against money rent. The Cossacks received two-thirds of the land, to be held in common, but in lieu of redemption payments they had to serve 20 years in the army. The serfs in mines and households were freed but received no economic assets.

Redemption payments, however, soon proved too burdensome, village restrictions were tight, and the allotment land area declined, all of which led to renewed restlessness and disturbances. Following the revolt of 1905, the government, under Pyotr Stolypin, tried to create middle-class, independent farmers by replacing the village tenure with private ownership, consolidating holdings, and encouraging land purchase by individuals; but the time was too short for effective implementation. The Soviet Revolution overthrew the tsarist regime and introduced the concepts of public ownership and collectivization.

By decree in 1918, the Soviets abolished private ownership of land, made farming the sole basis of landholding, and declared collectivization a major objective of policy. Marketing of agricultural products became a state monopoly. In 1929 Stalin embarked on a full course of collectivization, and by 1938 collective farms occupied 85.6 percent of the land and state farms 9.1 percent. Credit facilities and tractor stations supplemented collectivization, while agricultural production was integrated in the national plan for industrialization and development.

The costs of Soviet reform included the destruction of capital and the death of large numbers of kulaks, or rich peasants. Total output and productivity increased, however, and capital formation was made possible through forced saving, taxes, and regulated prices. The peasant received extensive social services such as health care, and education and better working conditions. The objectives of the decree of 1918 have been fully realized.

Reform in eastern Europe

Reform in eastern Europe was complicated by the fact that most of the eastern European countries remained under foreign rule until the middle of the 19th century or later. In Hungary, the Decree of 1853 abolished the *robot,* or forced labour and feudal dues, freed the serfs, liberalized land transaction, and encouraged consolidation. The Romanian reform of 1864 freed the serfs and distributed both the land and the redemption payments in proportion to the number of cows or oxen each peasant had. Formal emancipation in Bulgaria was introduced by the Turkish government in the 1850s, but actual reform came in 1880, after independence. Each peasant, including sharecroppers and wage workers, who had worked the land for 10 years without interruption, was entitled to the land he had cultivated. With the exception of Bulgaria, the distribution of ownership throughout most of eastern Europe remained highly uneven. Political instability reached a dangerous point between the two world wars. Following World War II, the eastern European countries established Communist governments with a strong tendency toward collective, cooperative, and mechanized agriculture.

Land reform in Mexico

Mexico. The Mexican reform of 1915 followed a revolution and dealt mainly with lands of Indian villages that had been illegally absorbed by neighbouring haciendas (plantations). Legally there was no serfdom; but the Indian wage workers, or peons, were reduced to virtual serfdom through indebtedness. Thus, the landlords were masters of the land and of the peons. The immediate aim of reform was to restore the land to its legal owners, settle the title, and use public land to reconstruct Indian villages. The motives were mainly to reduce poverty and inequality and to secure political stability, which was then in the balance. A decree of 1915 voided all land alienations that had taken place illegally since 1856 and provided for extracting land from haciendas to reestablish the collective Indian villages, or ejidos. The 1917 constitution reaffirmed those provisions but also guaranteed protection of private property, including haciendas. Nevertheless, a combination of loopholes, litigation, and reactionary forces slowed implementation, and effective reform came only after passage of the Agrarian Code of 1934 and the sympathetic efforts of Pres. Lázaro Cárdenas.

The reform restored many villages and freed the peons, but land concentration and poverty continued. In 1950, more than 31 percent of the private cropland was owned by fewer than 0.5 percent of the owners. Small-scale operation was retained or encouraged, a fact explaining the decline of output in the early years. More recently, efficiently run farms have been exempted from distribution.

The social and political impact was more positive. The peasants acquired more land and liberty, and control by landlords was reduced, although it was replaced by village restrictions. At least legally, farming became the basis of landholding. Some have seen in land reform the reason for Mexico's political stability, although there have been sporadic peasant uprisings and other violent encounters.

REFORMS SINCE WORLD WAR II

Communist and capitalist reforms

Recent decades have witnessed widespread, comprehensive reform programs, but the concept has undergone major changes. The eastern European countries and China originally followed the Soviet model, with different modifications in the individual countries. A few other countries have continued to follow that model, with major emphasis on "land to the tiller," cooperation, collective ownership, large-scale operation, and mechanization, and with economic development as the common denominator. In capitalist-oriented reforms, private ownership, family farming, and dual tenures have remained basic objectives with the aim of promoting democracy, equality, stability, and development. Under the influence and with the guidance of the United Nations, nonsocialist reforms of the 1950s were equated with community development and emphasized institutional and rural self-help in addition to land redistribution. In the 1960s the emphasis shifted to agricultural productivity and economic development by means of large-scale operation, new technology, and cooperation. The 1970s witnessed the advent of "integrated rural development" as the focus of reform and as a way of combining productive activities with improvements in the social and physical infrastructure. The integrated approach, however, soon proved to be unmanageable, and the emphasis shifted to the "target" group as the focus of reform. The most recent conception of reform has been to satisfy "basic needs," with or without land distribution, although no policymaker in the capitalist countries would openly question the idea of land redistribution or the creation of small family farms. These experiments with the concept of reform have been accompanied by attempts to broaden the concept to incorporate women as equal beneficiaries of reform in their own right. The results have been mixed.

Japan. The Japanese reform came immediately after World War II at the insistence of the Allied Occupation Army. The reform was designed to fit the uniquely high literacy rate and advanced industrial level of the country. Although the Meiji government had formally abolished feudalism and declared the land to be the property of the peasants, usurpation of land by the rich and by moneylenders had created classes of perpetual tenants and absentee landlords. In 1943, 66 percent of the land was operated by tenants against rent in kind that averaged 48 percent of the farmers' product, while population pressure resulted in fragmentation of holdings. The social class structure was closely tied to tenure, the owners in each village being at the top of the structure. Conflict between landlords and peasants was widespread.

After the war, the crisis was revived by food shortages, the breakdown of the urban economy, and the return of absentee landlords to the land. The Occupation Army insisted on reform, presumably to democratize the society and rehabilitate the economy. The reform law of 1946 established a ceiling on individual holdings and provided for expropriation and resale of excess land to the tenants against long-term payments. The government compensated the landlords in cash and bonds redeemable in 30 years. Tenants were protected by contract, and rents were reduced to a maximum of 25 percent of the product. The redistributed land was made inalienable, though this restriction was relaxed four years later. The program also provided for marketing and credit cooperatives. An important supplementary measure was the Local Autonomy Law of 1947, which decentralized the power structure and put village affairs in the hands of the villagers.

Within two years tenancy declined by more than 80 percent. Rent control and land distribution helped to equalize incomes in the villages and rehabilitate the sociopolitical status of the peasants. Crop yields per unit of land increased, but despite improved techniques the output per worker declined. In general the reform seemed to realize the objectives of the reformers and the peasants, although smallness of scale, low per capita incomes, underemployment, and insufficient mechanization have persisted. Even black market rents developed. These problems were tolerable because their effects were mitigated by the upsurge of the urban economy and the ability of the Japanese farmer to supplement the family income from nonagricultural employment. Even so, the farmers continue to depend on government subsidy to stay in farming.

Egypt. The Egyptian reform of 1952 followed the revolution that overthrew the monarchy and brought young middle-class leaders to the helm. Though affecting only about 12 percent of the arable land, it was applied thoroughly and touched all aspects of rural life. Egypt had two main forms of tenure: private ownership and *waqf,* or land held in trust and dedicated to charitable or educational purposes. *Waqf* land was inalienable, but private land was subject to speculation and concentration. In 1950, 1 percent of the owners had more than 20 percent of the private land, and 7 percent had more than two-thirds. The operating unit was small, with 77 percent of all the holdings occupying less than one acre each. Tenancy was widespread and rents were exorbitant. The peasants were exploited by middlemen who sublet the land to tenants, mediated between them and the market, and extended credit at high rates of interest.

The revolutionary reformers aimed at abolishing feudalism, recruiting peasant support, promoting economic development, and bringing the villagers back into the stream of national life. The Agrarian Reform Law of 1952 put a ceiling on individual holdings at 200 *faddān*s (one *faddān* = 1.038 acres), later reduced to 100 *faddān*s, with special allowance for male children. The excess land was expropriated and distributed to the peasants in parcels not exceeding five *faddān*s. Compensation was given in bonds, while land recipients had to repay in annual installments. The new owners were obligated to join cooperatives for production, marketing, and credit. Tenancy conditions were also regulated, with contract replacing traditional terms; rent could not exceed 50 percent of the product, nor could a tenant hold more than 50 acres, to avoid subletting. An interesting feature of the reform was the special attention given to college graduates by allowing them up to 20-*faddān* parcels.

The Reform Law of 1952

The reform was enforced quickly and had a great impact on the morale of the peasants. The economic effects, however, were minor since agriculture was intensive and land yield high. Producer cooperatives served only to offset the impact of distribution on the scale of operation. Some increases in yield have been claimed, but the evidence is still inconclusive. Furthermore, little capital was redirected into productive investment since the compensation bonds were not negotiable. Peasant savings remained limited, income increments being spent mostly on consumption. Finally, underemployment in agriculture has remained widespread. The defects of the agrarian structure continue to prevail, and relatively large ownerships exist, while certain groups in Egypt are calling for reversal of the reform.

The social and political effects, however, were far reaching. Redistribution and regulation of rent raised the incomes of small owners and tenants. Cooperatives replaced the middleman and captured his share for the farmer. The peasant gained social status and enjoyed a higher level of political participation, mostly in support of the revolutionary regime. These effects, however, can be easily exaggerated. The peasants became dependent on the cooperatives whether they liked them or not. Great differences in landholding continued to exist, and peasant incomes remained low. Black market rents appeared. The example of Egypt suggests that successful reform in densely populated countries requires an upsurge in the industrial sector to relieve population pressure and permit technical advance and higher productivity in agriculture.

Southeast Asia. The model of Japan's reform has been attempted in Southeast Asia, especially in Taiwan, South Korea, and South Vietnam, all influenced by American experts and by the anti-Communism of the respective governments. The objectives were to sustain the political order, raise living standards, and promote some degree of economic development. The reforms began with regulation of tenancy, restriction of rent, and the institution of written contract for leases, following which tenants were to be transformed into owners. Taiwan's reform was implemented between 1949 and 1953, in three stages. First, rents, which had sometimes reached 70 percent of the product, were reduced to 37.5 percent. Next, tenant-farmed public land was sold to the tenants. Finally, tenant-farmed private land was bought by the government and resold to the tenants.

The Vietnamese reform was introduced in 1955. Rents were reduced to a maximum of 25 percent of the product. A ceiling of 247 acres (100 hectares) was put on individual holdings, however, and only the excess land was subject to redistribution in parcels of 7.4 to 12.4 acres (three to five hectares) to the tenants. The collapse of the South Vietnamese regime and the unification of South and North Vietnam ended that reform and replaced it with the socialist model of North Vietnam.

The reform in Taiwan, as in South Vietnam prior to unification, was supplemented by other measures described as community development, such as adult education, credit facilities, improved technology, and other social services. Though land consolidation was attempted, the scale of operation was little affected. The main effect seems to have been the regulation of tenancy and the redistribution of rent incomes. An innovation of Taiwan's reform was the partial compensation of landlords with industrial shares in public enterprises, which helped them and helped industry.

Taiwan's reform has been hailed as a major success, in both economic and political terms. Some observers, however, are unwilling to reach such a conclusion until restrictions are removed and the peasants have a free choice of tenure and farm organization.

South Korea's land reform (under the Land Reform Law of June 1949) roughly followed the Japanese model by removing tenancy, creating small ownerships, implementing the law thoroughly and promptly, and depending heavily on nonagricultural (basically industrial) employment to absorb labour and supplement rural income. Like the Japanese and Taiwanese reforms, Korea's successful reform was generously supported by foreign aid.

The Philippines introduced a reform program in 1963, which aimed primarily at replacing share tenancy with lease contracts and eventually with ownership, and at revitalizing agriculture through extension services. By the mid-1980s the program had given titles to about 400,000 tenants and secure leases to another 600,000, but the economic viability of the new units has been uncertain because of their small scale and the lack of supplementary facilities. The main effects initially were seed improvement, greater use of fertilizers, and an increase in contractual tenancy. To combat the negative effects of small-scale farming, the Philippine government has resorted to what it calls the "compact farm," which is a voluntary grouping of small farms to be operated under one management as one consolidated farm. The problem of surplus labour, however, remains to be solved.

Various other reforms have been introduced in Southeast Asia, but the only innovative program has been that of Malaysia. The program in Malaysia has been highly organized and development oriented. It tries to promote social and economic objectives by emphasizing the production of rubber and palm oil for export and gradually transforming the landless into hereditary tenants on newly reclaimed and settled plantations. A typical plantation covers 4,500 to 5,000 acres (1,800 to 2,000 hectares) of jungle land and absorbs about 400 families. The land is cleared and planted by contract, and a village is constructed, with all the necessary services, before the settlers arrive. Each house has a quarter of an acre for a household garden. Cropland is divided in blocks of 120 to 200 acres (48 to 80 hectares), to be worked by a team of 15 to 25 people until the plants have matured. Upon maturation, each settler receives a share by lottery and a lease title for 99 years. This tenure arrangement precludes alienation, subdivision, or subleasing; it thus protects the tenant farmer and sidesteps the Islāmic laws of inheritance, which tend toward fragmentation of the land.

Reform in Malaysia

The settler is responsible for the cost of clearing and planting, but the government pays the administrative costs. The settler is guaranteed supplementary employment to earn subsistence income pending maturity of the plants, and cultivation is guided by experts. The rate of settlement is determined by the overall economic plan. It is clear that landholding has become tied to cultivation; fragmentation and diseconomies of scale have been avoided, and cultivation has become a rational economic operation. The Malaysian program has much in common with the cooperative settlements of Israel and the Gezira Scheme in The Sudan.

Latin America. Except for the early example of Mexico, reform in Latin America has been recent and appears to have come only in response to the threat of social and political instability and mounting international pressures. Reform in Latin America after World War II must be seen against a background of rapidly increasing population and of extreme contrasts between plantation economies and small units; high concentration of land ownership, income, and power and dire poverty; modern farming and relatively backward cultivation methods; and nationalism and extensive foreign ownership of land. In addition, Latin-American society is complicated by its ethnic mixtures and by dependence on staple trade items such as sugar, tobacco, cocoa, coffee, and beef cattle.

Reform in Latin America has reflected the ideologies and objectives of the regime in power. Brazil has had several attempts at reform. The measures have been indirect and relatively mild, the most important being taxation of idle land and large plantations and reclamation and settlement of the Amazon region, with provisions for credit and tenancy protection. The results have been modest, however, largely because of the physical and biological hardships faced by settlers in the tropical Amazon environment. Peru has deviated by creating collective administrations of the nationalized feudal estates. The title resides in the nation, and the estates are run by the Agricultural Societies of Social Interest (SAIS), a mechanism devised to avoid breaking up economically efficient enterprises rather than to modify the tenure institutions.

Reform in Cuba

At the other end of the Latin-American spectrum is the Cuban reform that followed the revolution of 1958. Cuba retained private ownership but reduced it substantially in favour of the public sector. As proclaimed a few months before the overthrow of the old regime, the reform aimed at the elimination of *latifundia* tenure, expropriation of land owned by foreign companies, higher standards of living for the peasantry, and national economic development. It began by setting a ceiling of 30 *caballerías* (one *caballería* = 33 acres, or 13.4 hectares) on individual holdings, with a maximum of 100 *caballerías* if economic operations required such a scale. All foreign-owned land was nationalized. Public land on which rice and cattle were raised was converted into state farms, and the peasants became permanent wage workers on these farms. Sugar plantations were converted into cooperatives to avoid their subdivision into small uneconomic units. Before long the ceiling on individual holdings was lowered to five *caballerías*, and all such holdings became private family farms. The rest were nationalized, and the expropriated owners were compensated with a pension for life. The reform was supplemented by the organization of national farmer associations; people's stores; credit, housing, and educational facilities; and the production of machinery and fertilizers. In 1963 a major reorganization of state farms took place; they were subdivided on the basis of crop specialization into smaller operational units of about 469 *caballerías*.

Effects of the reform were comprehensive and immediate. The tenure institutions were radically changed in favour of public ownership, while *minifundia* and tenancies were abolished. Socially and politically, the reform realized the objectives of the reformers. Economically, the government claimed higher yields of sugarcane, vegetables, and fruit, but this claim has been disputed by foreign observers.

Other Latin-American reforms fall between those of Brazil and Cuba, though closer to the former than to the latter in comprehensiveness and thoroughness. For example, the reform in Costa Rica has overlooked land concentration and income inequality and concentrated on the squatters, or *parásitos,* who in 1961 numbered between 12,000 and 16,000 people. The reform aimed at legalizing existing squatter holdings, preventing further squatting, and conserving virgin land. Even this modest program was implemented very slowly. As late as 1973, 7.3 percent of the landholdings comprised 67 percent of the total agricultural land. Colombia has had reform programs for at least 30 years, but concentration of ownership, fragmented holdings, backward methods of cultivation, inequality of income distribution, and widespread poverty have remained characteristic; in 1970, 4.3 percent of the holdings contained 67.4 percent of the total area.

Chile

Chile undertook various reform programs before achieving concrete results. In 1962 a program was enacted to encourage settlement of new land, but only about 1,000 families were settled. A comprehensive reform was introduced in 1965 with three main objectives: to make the agricultural workers owners of the land they had cultivated previously, to increase agricultural and livestock production, and to facilitate social mobility and peasant participation in political life. The Chilean reform was unique in its method of implementation. Once the plantation had been designated for expropriation and the prospective owners selected, they were organized into *asentamientos,* or settlement groups. The group elected a committee to take charge of settlement. The members cultivated the land as a team for three to five years. Meanwhile they received training and guidance in social participation, decision making, and modern farming. Upon completion of the transition period, the land was divided among those who had shown promise, to be held outright and without restriction. All new owners were obligated to join cooperatives, the form of these being determined by the members. The socialist regime that came into office in 1970 expedited the expropriation process and the creation of settlement groups or cooperative farms under peasant committees. By 1972 all the potential land, which had been in farms larger than 200 acres (80 hectares), had been expropriated and reallocated. The new regime that took over in 1973 decided, however, to privatize the land and reverse much of the reform by returning large areas to the former owners, dissolving the cooperatives, and creating private ownerships in their place. Most of the reverse changes had been completed by 1979. Nevertheless, most of the excess land in farms of more than 200 acres remained in the hands of the reform beneficiaries. Owners of less than 12 acres (five hectares) were hardly affected; those who owned between 12 and 50 acres (five and 20 hectares) benefitted most. In the final analysis, less than 15 percent of the agricultural land was affected by the reform between 1965 and 1979 under three regimes.

Observers of the Latin-American scene have been pessimistic regarding the adequacy of these land reform programs. With the exception of Cuba, capital formation in agriculture has not increased substantially; the pattern of land distribution has undergone little change; social and political stability have remained in question; and the agrarian structure is still considered defective.

Other recent reforms. Attempts to reform the agrarian structure have been made in most other countries, with varying degrees of seriousness. India and Pakistan have concentrated on abolishing intermediaries who prevailed as survivals of traditional and feudal tenures. In India the tenants have become hereditary holders, with the title vested in the state. India has left reform to the states and emphasized peaceful and compensatory methods; hence the results have varied from one state to another. Pakistan, following the revolution of 1958, enacted a reform that made most of the tenants owners. In both countries, however, small-scale farming has persisted, while Pakistan has continued to tolerate and protect owners of up to 500 acres (200 hectares). In neither country has fragmentation been effectively reduced or have capital formation and cultivation methods significantly advanced.

India and Pakistan

In contrast, after the Communists came to power in China, private ownership was eliminated and the peasants were organized in village communes. Extensive supplementary measures have been tried, and the role and organization of the commune have varied according to the pressures on the economy. The most recent innovation in China's agriculture has been the "production responsibility system," which allows the commune to contract with its members for quotas of output; the members are free to sell the surplus on the open market. The change is seen as an incentive generator, but land cannot be rented, bought, sold, or used except as authorized by the commune. The effects of China's agrarian policy on peasant living conditions and the Chinese economy have been generally accepted as positive, genuine, and impressive.

In 1962 Iran made owners of most of the former sharecroppers, in the classic tradition of Western-type reform, mainly to create political stability. Given Iran's revolution of 1979, however, the reform evidently was not sufficient to sustain the old social order. Reform was also introduced in Syria, Iraq, Algeria, Libya, and other countries of the Middle East and North Africa following independence or revolution. Most of these reforms were influenced by the Egyptian example, with the state playing a major role. In all cases emphasis has been placed on farm cooperatives, although they have been largely ineffective.

In contrast, tropical Africa has witnessed a wave of innovative reform in recent years. Reform has sometimes come in "packages," which combine tenure reform and other measures affecting cultivation and productivity. Among the innovations is the "villagization," or *ujamaa,* program of Tanzania, according to which a group of families lives, works, and makes decisions together and shares the costs and benefits of farming the land. The program began as a voluntary movement in 1967, but by 1977 it had become almost mandatory. At the same time, "block farming" and individual holdings had become acceptable forms of cooperation. The Ujamaa Villages Act of 1975 made the village the main rural administration and development unit. The most radical reforms in Africa, however, have been those of Ethiopia in 1975 and of Mozambique in 1979. Both vested the land title in the nation and abolished rent, sale, and absentee control of the land. The land was placed in the hands of the tillers, who have guaranteed right of use for themselves and for their descendants. Except in the public sector, farming is a small, family operation with a high degree of equality of landholding but of uncertain efficiency.

CONCLUSIONS

Land reform and agrarian reforms have become synonymous, indicating that reform programs have become more comprehensive and encompass much more than the reform of land tenure or land distribution. Reform movements have recurred throughout history, as have the crises they are intended to deal with, because reform has rarely dealt with the roots of the crises. Reform has served as a problem-solving mechanism and therefore has only been extensive enough to cope with the immediate crisis. Reformers have often faced hard choices: to promote and sustain private ownership with inequality or to institute public or collective ownership with equality but with restrictions on the individuals' private interests; to spread employment by supporting labour-intensive, low-productivity techniques or to promote high productivity through capital-intensive, efficient methods; to pursue gradual "repair and maintenance" reform that is basically ineffective or to promote revolutionary, comprehensive, effective but disruptive reform. In capitalist reforms these contradictions have usually been resolved in favour of the first set of options; in socialist reforms, in favour of the second. Land tenure reform seems to have been of little significance in creating substantive economic change, although it has been important for improving the status of peasants and maintaining social and political stability. Most reforms have narrowed the gap between reform beneficiaries and other farmers through land redistribution and tenancy control, but only the comprehensive socialist reforms have narrowed the gap between agriculture and other sectors of the economy.

Land redistribution programs have had limited success for several reasons. They often have deprived the farm of the former landlord's contributions without providing a substitute. They have inhibited mobility of labour by giving the peasant a stake in the land, though only in the form of an inefficient minifarm. They frequently have threatened large, efficiently run farms and therefore have had to be compromised. They have provided compensation for the expropriated land and hence left wealth and income distribution largely unaffected. They have been conditional upon peasant participation in social and political activity and cooperative organization, even though the peasant was unprepared for these activities. Moreover, the redistribution of land has rarely been fortified by protective measures that could prevent reconcentration of ownership and the recurrence of crises. Nevertheless, major efforts have been expended by the Food and Agriculture Organization of the United Nations and other international bodies and by governments to devise viable frameworks for solving agricultural and rural problems emanating from defective agrarian structures.

BIBLIOGRAPHY. The philosophy and logic of land reform are presented in A. WHITNEY GRISWOLD, *Farming and Democracy* (1948, reissued 1963), a lucid scholarly analysis of the political significance of various farm structures and a discussion of the family farm as a cornerstone of democracy; FOLKE DOVRING, *Land and Labor in Europe in the Twentieth Century*, 3rd rev. ed. (1965), an examination of the political and ideological bases of land policy since the French Revolution; ELIAS H. TUMA, *Twenty-Six Centuries of Agrarian Reform: A Comparative Analysis* (1965), a synthesis of theory and history, with an attempt to formulate a general theory of agrarian reform, and "Agrarian Reform in Historical Perspective Revisited," *Comparative Studies in Society and History*, 21:3–29 (1979), which updates and retests the author's earlier theory; ERICH H. JACOBY, *Evaluation of Agrarian Structures and Agrarian Reform Programs* (1966), a checklist of administrative and organizational changes that contribute to reform success; and DOREEN WARRINER, *Land Reform in Principle and Practice* (1969), a provocative discussion of the evolution of reform and conflict between theory and practice.

More recent works discussing the same issues include WORLD CONFERENCE ON AGRARIAN REFORM AND RURAL DEVELOPMENT, *The Peasants' Charter* (1981), a declaration of principles and program of action by most of the nonsocialist developing countries; JOHN D. MONTGOMERY (ed.), *International Dimensions of Land Reform* (1984), an overview of how such issues are handled by international agencies; INGRID PALMER, *The Impact of Agrarian Reform on Women* (1985), an examination of fertility, health, and nutrition; JOHN P. POWELSON, *The Story of Land: A World History of Land Tenure and Agrarian Reform* (1988), a good starting place for study, with a comprehensive bibliography; DEMETRIOS CHRISTODOULOU, *The Unpromised Land: Agrarian Reform and Conflict Worldwide* (1990), an excellent resource list and a fine study of the circumstances which render land reform necessary; M. RIAD EL-GHONEMY, *The Political Economy of Rural Poverty: The Case for Land Reform* (1990), a convincing argument for reduced land concentration; and SOLON L. BARRACLOUGH, *An End to Hunger?: The Social Origins of Food Strategies* (1991), a study placing land reform issues for developing countries in the broader world context.

Reforms in the ancient world are addressed in W. WARDE-FOWLER, "Notes on Gaius Gracchus," *English Historical Review*, 20:209–227 and 417–433 (1905); E.G. HARDY, "Were the Lex Thoria of 118 B.C. and the Lex Agraria of 111 B.C. Reactionary Laws?," *Journal of Philosophy*, 31:268–286 (1910); IVAN M. LINFORTH, *Solon the Athenian* (1919, reprinted 1971); and W.J. WOODHOUSE, *Solon the Liberator: A Study of the Agrarian Problems in Attica in the Seventh Century* (1938, reprinted 1965).

ELIAS H. TUMA, *European Economic History: Tenth Century to the Present* (1971), summarizes the agrarian reform history of most countries of Europe since the Middle Ages. Early French reforms are examined by MARC BLOCH, *French Rural History: An Essay on Its Basic Characteristics* (1966; originally published in French, 1931); and GEORGES LEFEBVRE, *The Coming of the French Revolution: 1789*, bicentennial ed. (1989; originally published in French, 1939). Reform in tsarist Russia is addressed in GEROID T. ROBINSON, *Rural Russia Under the Old Régime: A History of the Landlord-Peasant World and a Prologue to the Peasant Revolution of 1917* (1932, reissued 1969).

Any study of recent decades must begin with *Progress in Land Reform: Sixth Report* (1976), issued jointly by the United Nations, its Food and Agriculture Organization (FAO), and the International Labour Organisation; and with two publications from the FAO, *Review and Analysis of Agrarian Reform and Rural Development in the Developing Countries Since the Mid-1960s* (1979?), and *Land Reform, Land Settlement, and Cooperatives* (semiannual), which offers feature articles, summaries of recent reform legislation, and a comprehensive bibliography of new literature.

Summaries and analyses of reform over broad regions can be found in JEAN LE COZ, *Les Réformes agraires: de Zapata à Mao Tsê-toung et la F.A.O.* (1974), which compares the Soviet model with the Chinese, and with reform in Latin America and the Middle East, and highlights the role of the FAO in reform implementation; V.E. STANIS, *Socialist Transformation of Agriculture: Theory and Practice* (1976; originally published in Russian, 1971), covering reform in eastern Europe after World War II; RUSSELL KING, *Land Reform: A World Survey* (1977), a good summary for the general reader, with references; DAVID A. PRESTON (ed.), *Environment, Society, and Rural Change in Latin America: The Past, Present, and Future in the Countryside* (1980), a collection of essays, pessimistic as to the efficacy of most reform; AJIT KUMAR GHOSE (ed.), *Agrarian Reform in Contemporary Developing Countries* (1983), covering developments in agrarian reform through the early 1980s, with illustrations from the field; M. RIAD EL-GHONEMY (ed.), *How Development Strategies Benefit the Rural Poor* (1984), an FAO summary of reform programs; and PETER UTTING, *Economic Reform and Third-World Socialism: A Political Economy of Food Policy in Transitional Societies* (1992), a lucid introduction to certain economic aspects of land reform in socialist economies.

Studies of specific countries and regions which also highlight important theoretical trends include MAURICE H. DOBB, *Soviet Economic Development Since 1917*, 6th ed. (1966); GEORGE YANEY, *The Urge to Mobilize: Agrarian Reform in Russia, 1861–1930* (1982); JOHN YIN, *Infrastructure of the Soviet Agriculture* (1991); ERIC J. HOOGLUND, *Land and Revolution in Iran, 1960–1980* (1982); ELIAS H. TUMA, *Economic and Political Change in the Middle East* (1987); ASGHAR SCHIRAZI, *Islamic Development Policy: The Agrarian Question in Iran* (1993); KENNETH M. LUNO, *The Pasha's Peasants: Land, Society, and Economy in Lower Egypt, 1740–1858* (1992), providing historical perspective; JAMES BROW and JOE WEERAMUNDA (eds.), *Agrarian Change in Sri Lanka* (1992); JAMES PUTZEL, *A Captive Land: The Politics of Agrarian Reform in the Philippines* (1992), a detailed and erudite study; JOHN O. HALEY and KOZO YAMAMURA (eds.), *Land Issues in Japan: A Policy Failure?* (1992); WILLIAM HINTON, *The Great Reversal: The Privatization of China, 1978–1989* (1990), discussing the mixture of collective and private approaches to land-use practices; VIVIENNE SHUE, *Peasant China in Transition: The Dynamics of Development Toward Socialism, 1949–1956* (1980), a close study of the initial stages of agricultural development and state control of the rural economy; ANTHONY Y.C. KOO, *Land Market Distortion and Tenure Reform* (1982), concentrating on Taiwan and Southeast Asia; LAURA J. ENRÍQUEZ, *Harvesting Change: Labor and Agrarian Reform in Nicaragua, 1979–1990* (1991), an unusual and particularly valuable work because of its revolutionary setting; STEVEN E. SANDERSON, *Agrarian Populism and the Mexican State: The Struggle for Land in Sonora* (1981), covering 1917–76, with much discussion of the national situation; and ROGER BARTRA, *Agrarian Structure and Political Power in Mexico* (1993), illuminating the debate over small-scale agriculture versus other, more corporate forms.

Some excellent sources on African land-reform issues and policies which have not been widely available until recently are ESSY M. LETSOALO, *Land Reform in South Africa: A Black Perspective* (1987), an essential majority perspective; FRED T. HENDRICKS, *The Pillars of Apartheid: Land Tenure, Rural Planning, and the Chieftaincy* (1990), which covers some of the homeland issues; MICHAEL DE KLERK (ed.), *A Harvest of Discontent: The Land Question in South Africa* (1991), which details the controversial South African situation; and THOMAS J. BASSETT and DONALD E. CRUMMEY (eds.), *Land in African Agrarian Systems* (1993), a collection of essays covering much of the continent.

(E.H.T.)

Language

Language interacts with every other aspect of human life in society, and it can be understood only if it is considered in relation to society. This article attempts to survey language (both spoken and written) in this light and to consider its various functions and the purposes it can and has been made to serve. Because each language is both a working system of communication in the period and in the community wherein it is used and also the product of its past history and the source of its future development, any account of language must consider it from both these points of view.

The science of language is known as linguistics. It includes what are generally distinguished as descriptive linguistics and historical linguistics. Linguistics is now a highly technical subject; it embraces, both descriptively and historically, such major divisions as phonetics, grammar, and semantics, dealing in detail with these various aspects of language. For a full account of the theory and methods of linguistic science, see the article LINGUISTICS. For a general survey of known living and dead languages, consult the article LANGUAGES OF THE WORLD.

This article is divided into the following sections:

Characteristics of language

DEFINITIONS OF LANGUAGE

Many definitions of language have been proposed. Henry Sweet, an English phonetician and language scholar, stated: "Language is the expression of ideas by means of speech-sounds combined into words. Words are combined into sentences, this combination answering to that of ideas into thoughts." The U.S. linguists Bernard Bloch and George L. Trager formulated the following definition: "A language is a system of arbitrary vocal symbols by means of which a social group cooperates." Any succinct definition of language makes a number of presuppositions and begs a number of questions. The first, for example, puts excessive weight on "thought," and the second uses "arbitrary" in a specialized, though legitimate, way (see below).

A number of considerations enter into a proper understanding of language as a subject:

1. Every physiologically and mentally normal person acquires in childhood the ability to make use, as both speaker and hearer, of a system of vocal communication that comprises a circumscribed set of noises resulting from movements of certain organs within his throat and mouth. By means of these he is able to impart information, to express feelings and emotions, to influence the activities of others, and to comport himself with varying degrees of friendliness or hostility toward persons who make use of substantially the same set of noises.

2. Different systems of vocal communication constitute different languages; the degree of difference needed to establish a different language cannot be stated exactly. No two people speak exactly alike; hence, one is able to recognize the voices of friends over the telephone and to keep distinct a number of unseen speakers in a radio broadcast. Yet, clearly, no one would say that they speak different languages. Generally, systems of vocal communication are recognized as different languages if they cannot be understood without specific learning by both parties, though the precise limits of mutual intelligibility are hard to draw and belong on a scale rather than on either side of a definite dividing line. Substantially different systems of communication that may impede but do not prevent mutual comprehension are called dialects of a language. In order to describe in detail the actual different speech patterns of individuals, the term idiolect, meaning the speech habits of a single person, has been coined.

Languages, dialects, and idiolects

3. Normally, people acquire a single language initially—their first language, or mother tongue, the language spoken by their parents or by those with whom they are brought up from infancy. Subsequent "second" languages are learned to different degrees of competence under various conditions, but the majority of the world's population remains largely monolingual. Complete mastery of two languages is designated as bilingualism; in a few special cases—such as upbringing by parents speaking different languages at home—speakers grow up as bilinguals, but ordinarily the learning, to any extent, of a second or other language is an activity superimposed on the prior

mastery of one's first language and is a different process intellectually.

4. Language, as described above, is species-specific to man. Other members of the animal kingdom have the ability to communicate, through vocal noises or by other means, but the most important single feature characterizing human language (that is, every individual language), against every known mode of animal communication, is its infinite productivity and creativity. Human beings are unrestricted in what they can talk about; no area of experience is accepted as necessarily incommunicable, though it may be necessary to adapt one's language in order to cope with new discoveries or new modes of thought.

Animal communication systems are by contrast very tightly circumscribed in what may be communicated. Indeed, displaced reference, the ability to communicate about things outside immediate temporal and spatial contiguity, which is fundamental to speech, is found elsewhere only in the so-called language of bees. Bees are able, by carrying out various conventionalized movements (referred to as bee dances) in or near the hive, to indicate to others the locations and strengths of nectar sources. But nectar sources are the only known theme of this communication system. Surprisingly, however, this system, nearest to human language in function, belongs to a species remote from man in the animal kingdom and is achieved by very different physiological activities from those involved in speech. On the other hand, the animal performance superficially most like human speech, the mimicry of parrots and of some other birds that have been kept in the company of humans, is wholly derivative and serves no independent communicative function. Man's nearest relatives among the primates, though possessing a vocal physiology very similar to that of humans, have not developed anything like a spoken language.

HISTORICAL ATTITUDES TOWARD LANGUAGE

As is evident from the introduction above, human life in its present form would be impossible and inconceivable without the use of language. People have long recognized the force and significance of language. Naming—applying a word to pick out and refer to a fellow human being, an animal, an object, or a class of such beings or objects—is only one part of the use of language, but it is an essential and prominent part. In many cultures men have seen in the ability to name an ability to control or to possess; this explains the reluctance, in several primitive and other communities, with which names are revealed to strangers and the taboo restrictions found in several parts of the world on using the names of persons recently dead. Lest it be thought that attitudes like this have died out in modern civilized communities, it is instructive to consider the widespread and perhaps universal taboos on naming directly things considered obscene, blasphemous, or very fearful. Indeed, use of euphemistic substitutes for words referring to death and to certain diseases actually seems to be increasing in some civilized areas.

The practice of naming

Not surprisingly, therefore, several independent traditions ascribe a divine or at least a supernatural origin to language or to the language of a particular community. The biblical account, representing ancient Jewish beliefs, of Adam's naming the creatures of the Earth under God's guidance is well known:

> So out of the ground the Lord God formed every beast of the field and every bird of the air, and brought them to the man to see what he would call them; and whatever the man called every living creature, that was its name (Gen. 2:19).

Norse mythology preserves a similar story of divine participation in the creation of language, and in India the god Indra is said to have invented articulate speech. In the much more sophisticated debate on the nature and origin of language given in Plato's Socratic dialogue *Cratylus*, Socrates is made to speak of the gods as those responsible for first fixing the names of things in the proper way.

A similar divine aura pervades early accounts of the origin of writing. The Norse god Odin was held responsible for the invention of the runic alphabet. The inspired stroke of genius whereby the ancient Greeks adapted a variety of the Phoenician consonantal script so as to represent the distinctive consonant and vowel sounds of Greek, thus producing the first alphabet such as is known today, was linked with the mythological figure Cadmus, who, coming from Phoenicia, was said to have founded Thebes and introduced writing into Greece. The Arabs had a traditional account of their script, together with the language itself, being given to Adam by God.

The later biblical tradition of the Tower of Babel (Gen. 11:1–9) exemplifies three aspects of early thought about language: (1) divine interest in and control over its use and development, (2) a recognition of the power it gives to man in relation to his environment, and (3) an explanation of linguistic diversity, of the fact that people in adjacent communities speak different and mutually unintelligible languages, together with a survey of the various speech communities of the world known at the time to the Hebrews.

Biblical traditions concerning language

The origin of language has never failed to provide a subject for speculation, and its inaccessibility adds to its fascination. Informed investigations of the probable conditions under which language might have originated and developed are seen in the late-18th-century essay of the German philosopher Johann Gottfried von Herder, "Abhandlung über den Ursprung der Sprache" ("Essay on the Origin of Language"), and in numerous other treatments. But people have tried to go further, to discover or to reconstruct something like the actual forms and structure of man's first language. This lies forever beyond the reach of science, in that spoken language in some form is almost certainly coeval with *Homo sapiens*. The earliest records of written language, the only linguistic fossils man can hope to have, go back no more than about 4,000 or 5,000 years. Attempts to derive human speech from imitations of the cries of animals and birds or from mere ejaculations of joy and grief, as if onomatopoeia were the essence of language, were ridiculed for their inadequacy by the Oxford philologist F. Max Müller in the 19th century and have been dubbed the bowwow and pooh-pooh theories.

On several occasions attempts have been made to identify one particular existing language as representing the original or oldest tongue of mankind, but, in fact, the universal process of linguistic change rules out any such hopes from the start. The Greek historian Herodotus told a story that King Psammetichus of Egypt caused a child to be brought up without ever hearing a word spoken in its presence. On one occasion it ran up to its guardian as he brought it some bread, calling out "bekos, bekos"; this, being said to be the Phrygian word for bread, proved that Phrygian was the oldest language of mankind. The naïveté and absurdity of such an account have not prevented its repetition elsewhere and at other times.

In Christian Europe the position of Hebrew as the language of the Old Testament gave valid grounds through many centuries for regarding Hebrew, the language in which God addressed Adam, as the parent language of all mankind. Such a view continued to be expressed even well into the 19th century. Only since the mid-1800s has linguistic science made sufficient progress finally to clarify the impracticability of speculation along these lines.

When people have begun to reflect on language, its relation to thinking becomes a central concern. Several cultures have independently viewed the main function of language as the expression of thought. Ancient Indian grammarians speak of the soul apprehending things with the intellect and inspiring the mind with a desire to speak; and in the Greek intellectual tradition Aristotle declared, "Speech is the representation of the experiences of the mind" (*On Interpretation*). Such an attitude passed into Latin theory and thence into medieval doctrine. Medieval grammarians envisaged three stages in the speaking process: things in the world exhibit properties; these properties are understood by the mind of man; and, in the manner in which they have been understood, so they are communicated to others by the resources of language.

Language and thinking

Rationalist writers on language in the 17th century gave essentially a similar account: speaking is expressing thoughts by signs invented for the purpose, and words of different classes (the different parts of speech) came into being to correspond to the different aspects of thinking.

Such a view of language continued to be accepted as generally adequate and gave rise to the sort of definition proposed by Henry Sweet and quoted above. The main objection to it is that it either gives so wide an interpretation to thought as virtually to empty the word of any specific content or gives such a narrow interpretation of language as to exclude a great deal of normal usage. A recognition of the part played by speaking and writing in social cooperation in everyday life has highlighted the many and varied functions of language in all cultures, apart from the functions strictly involved in the communication of thought, which had been the main focus of attention for those who approached language from the standpoint of the philosopher. To allow for the full range of language used by speakers, more comprehensive definitions of language have been proposed in recent years on the lines of the second one quoted above (*i.e.*, "A language is a system of arbitrary vocal symbols by means of which a social group cooperates").

A rather different criticism of accepted views on language began to be made in the 18th century, most notably by the French philosopher Étienne Bonnot de Condillac in "Essai sur l'origine des connaissances humaines" (1746; "Essay on the Origin of Human Knowledge") and by Johann Gottfried von Herder. These men were concerned with the origin and development of language in relation to thought in a way that earlier students had not been. The medieval and rationalist views implied that man as a rational, thinking creature invented language to express his thoughts, fitting words to an already developed structure of intellectual competence. With the examination of the actual and the probable historical relations between thinking and speaking, it became more plausible to say that language emerged not as the means of expressing already formulated judgments, questions, and the like but as the means of thought itself, and that man's rationality developed together with the development of his capacity for speaking.

Language as a means of thought itself

The relations between thought and speech are certainly not fully explained today, and it is clear that it is a great oversimplification to define thought as subvocal speech, in the manner of some behaviourists. But it is no less clear that propositions and other alleged logical structures cannot be wholly separated from the language structures said to express them. Even the symbolizations of modern formal logic are ultimately derived from statements made in some natural language and are interpreted in that light.

The intimate connection between language and thought, as opposed to the earlier assumed unilateral dependence of language on thought, opened the way to a recognition of the possibility that different language structures might in part favour or even determine different ways of understanding and thinking about the world. Obviously, all people inhabit a broadly similar world, or they would be unable to translate from one language to another; but, equally obviously, they do not all inhabit a world exactly the same in all particulars, and translation is not merely a matter of substituting different but equivalent labels for the contents of the same inventory. From this stem the notorious difficulties in translation, especially when the systematizations of science, law, morals, social structure, and so on are involved. The extent of the interdependence of language and thought—linguistic relativity, as it has been termed—is still a matter of debate, but the fact of such interdependence can hardly fail to be acknowledged.

WAYS OF STUDYING LANGUAGE

Languages are immensely complicated structures. One soon realizes how complicated any language is when trying to learn it as a second language. If one tries to frame an exhaustive description of all the rules embodied in one's language—the rules by means of which a native speaker is able to produce and to understand an infinite number of correct, well-formed sentences—one can easily appreciate the complexity of the knowledge acquired by a child in mastering his mother tongue. The descriptions of languages written so far are in most cases excellent as far as they go, but they still omit more than they contain of an explicit account of a native speaker's competence in his language, by virtue of which one calls him a speaker of English, French, Swedish, or Swahili. The most recent developments in the study of language have served to reveal just how much more there is to do to bring palpable fact within systematic statement.

A detailed treatment of the science of linguistics is found elsewhere (see LINGUISTICS). Here it is proposed simply to give a brief outline of the way language or languages can be considered and described from different points of view, or at different levels, each contributing something essential and unique to a full understanding of the subject.

Phonetics and phonology. The most obvious aspect of language is speech. Speech is not essential to the definition of an infinitely productive communication system, such as is constituted by a language. But, in fact, speech is the universal material of human language, and the conditions of speaking and hearing have, throughout human history, shaped and determined its development. The study of speech sounds and of the physiology of speaking is called phonetics; this subject is dealt with further below, as well as in the article SPEECH: *The phonetics of speech.* Articulatory phonetics relates to the physiology of speech and acoustic phonetics to the physics of sound waves, their transmission and reception.

Speech, the universal material of human language

Phonetics covers much of the ground loosely referred to in language study as pronunciation. But, from a rather different point of view, speech sounds are also studied in phonology. Every language makes use of a very wide range of the articulations and resultant sounds that are available within the human vocal and auditory resources. Each language uses a somewhat different range, and this is partly responsible for the difficulty of learning to speak a foreign language and for speaking it "with an accent." But mere repertoires of sounds are not all that is involved. Far fewer general classes of sounds are distinctive (carry meaning differences) in any language than the number of sounds that are actually phonetically different. The English *t* sounds at the beginning and end of "tot" and in the two places in "stouter" are all different, though these differences are not readily noticed by English speakers; and, rightly, the same letter is used for them all. Similar statements could be made about most or all of the other consonant and vowel sounds in English.

What is distinctive in one language may not be distinctive in another or may be used in a different way; this is an additional difficulty to be overcome in learning to speak and understand a foreign language. In Chinese and in several other languages loosely called tone languages, the pitch, or tone, on which a syllable is said helps to distinguish one word from another: *ma* in northern Chinese on a level tone means "mother," on a rising tone means "hemp," and on a falling tone means "to curse." In English and in most of the languages of Europe (though not all—Swedish and Norwegian are exceptions) pitch differences do not distinguish one word from another, but form part of the intonation tunes that contribute to the structure and structural meaning of spoken sentences.

Languages differ in the ways in which consonant and vowel sounds can be grouped into syllables in words. English and German tolerate several consonants before and after a single vowel: "strengths" has three consonant sounds before and three after a single vowel sound (*ng* and *th* stand for one sound each). Italian does not have such complex syllables, and in Japanese and Swahili, for example, the ratio of consonant and vowel sounds in syllables and in words is much more even. Speakers of such languages find English words of the sort just mentioned very hard to pronounce, though to an Englishman they are perfectly "natural," "natural" in this context meaning "within the sounds and sound sequences whose mastery is acquired in early childhood as part of one's mother tongue."

All these considerations relating to the use of speech sounds in particular languages fall under the general heading of phonology; phonology is often regarded as one component of language structure.

Grammar. The other component is grammar. There is more to language than sounds, and words are not to be regarded as merely sequences of syllables. The concept of the word is a grammatical concept; in speech, words are not separated by pauses, but they are recognized as recurrent units that make up sentences. Very generally, grammar is concerned with the relations between words in sentences. Classes of words, or parts of speech, as they are often called, are distinguished because they occupy different places in sentence structure, and in most languages some of them appear in different forms according to their function (English "man," "men"; "walk," "walked"; "I," "me"; and so on). Languages differ in the extent to which word-form variation is used in their grammar; Classical Chinese had almost none, English does not have much, and Latin and Greek had quite a lot. Conversely, English makes much more use of word order in grammar than did Latin or Greek.

Relations between words in sentences

Traditionally, grammar has been divided into syntax and morphology, syntax dealing with the relations between words in sentence structure, morphology with the internal grammatical structure of words. The relation between "boy" and "boys" and the relationship (irregular) between "man" and "men" would be part of morphology; the relation of concord between "the boy [or "man"] is here" and "the boys [or "men"] are here" would be part of syntax. It must, however, be emphasized that the distinction between the two is not as clear-cut as this brief illustration might suggest. This is a matter for debate among linguists of different persuasions; some would deny the relevance of distinguishing morphology from syntax at all, referring to grammatical structure as a whole under the term syntax.

Grammar is different from phonology, though the word grammar is often used comprehensively to cover both aspects of language structure. Categories such as plural, past tense, and genitive case are not phonological categories. In spoken language they are, like everything else, expressed in speech sounds, but within a language these may be very different for one and the same category. In English noun plurals, the added *-s* in "cats," the vowel changes in "man, men" and in "goose, geese," and the *-en* in "oxen" are quite different phonologically; so are the past-tense formatives such as *-ed* in "guarded," *-t* in "burnt," vowel change in "take, took," and vowel and consonant change in "bring, brought." In Latin the genitive case can be represented in singular nouns by *-ī, -is, -ae, -ūs*, and *-eī*. The phonological difference does not matter, provided only that the category distinction is somehow expressed.

The same is true of the orthographic representation of grammatical differences, and the examples just given illustrate both cases. This is why the grammar of written language can be dealt with separately. In the case of dead languages, known with certainty only in their written forms, this must necessarily be done; insofar as the somewhat different grammar of their spoken forms made use of sound features not represented in writing (*e.g.*, stress differences), this can, at best, only be inferred or reconstructed.

Meaning expressed by grammatical forms, vocabulary, and intonation

Grammatical forms and grammatical structures are part of the communicative apparatus of languages, and along with vocabulary, or lexicon (the stock of individual words in a language), they serve to express all the meanings required. Spoken language has, in addition, resources such as emphatic stressing and intonation (see below). This is not to say, however, that grammatical categories can be everywhere directly related to specific meanings. Plural and past tense are fairly clear as regards meaning in English, but even here there are difficulties; in "if I knew his address I would tell you," the past-tense form "knew" refers not to the past but to an unfulfilled condition in the present. In some other languages greater problems arise. The gender distinctions of French, German, and Latin are very much part of the grammar of these languages, but only in a small number of words do masculine, feminine, and neuter genders correspond with differences of sex, or with any other category of meaning in relation to the external world (see also LINGUISTICS).

Semantics. Language exists to be meaningful; the study of meaning, both in general theoretical terms and in reference to a specific language, is known as semantics. It embraces the meaningful functions of phonological features, such as intonation, and of grammatical structures and the meanings of individual words. Once again, it must be stressed that questions arising from the relations between grammar and meaning and between grammar and phonology are the subjects of continuing controversy today (see also LINGUISTICS: *Semantics*).

Language variants

The word language contains a multiplicity of different designations. Two senses have already been distinguished: language as a universal species-specific capability of mankind, and languages as the various manifestations of that capability, as with English, French, Latin, Swahili, Malay, and so on. There is, of course, no observable universal language over and above the various languages that have been or are spoken or written; but one may choose to concentrate on the general and even the universal features, characteristics, and components of different languages and on the ways in which the same sets of descriptive procedures and explanatory theories may be applied to different languages. In so doing one may refer to language (in general) as one's object of study. This is what is done by linguists, or linguistic scientists, persons devoting themselves to the scientific study of languages (as opposed to the popular sense of polyglots, persons having a command of several different languages).

DIALECTS

It has already been pointed out that no two persons speak exactly alike, and within the area of all but the smallest speech communities (groups of people speaking the same language) there are subdivisions of recognizably different types of language, called dialects, that do not, however, render intercommunication impossible nor markedly difficult. Because intercomprehensibility lies along a scale, the degree required for two or more forms of speech to qualify as dialects of a single language, instead of being regarded as separate languages, is not easy to quantify or to lay down in advance, and the actual cutoff point must in the last resort be arbitrary. In practice, however, the terms dialect and language can be used with reasonable agreement. One speaks of different dialects of English (Southern British English, Northern British English, Scottish English, Midwest American English, New England American English, Australian English, and so on, with, of course, many more delicately distinguished subdialects within these very general categories), but no one would speak of Welsh and English or of Irish and English as dialects of a single language, although they are spoken within the same areas and often by people living in the same villages as each other. (Ro.H.R./Ed.)

Intercomprehensibility of dialects

Varieties of dialects. *Geographic dialects.* The most widespread type of dialectal differentiation is geographic. As a rule, the speech of one locality differs at least slightly from that of any other place. Differences between neighbouring local dialects are usually small, but, in travelling farther in the same direction, differences accumulate. Every dialectal feature has its own boundary line, called an isogloss (or sometimes heterogloss). Isoglosses of various linguistic phenomena rarely coincide completely, and by crossing and interweaving they constitute intricate patterns on dialect maps. Frequently, however, several isoglosses are grouped approximately together into a bundle of isoglosses. This grouping is caused either by geographic obstacles that arrest the diffusion of a number of innovations along the same line or by historical circumstances, such as political borders of long standing, or by migrations that have brought into contact two populations whose dialects were developed in noncontiguous areas.

Local and regional dialects

Geographic dialects include local ones (*e.g.*, the Yankee English of Cape Cod or of Boston, the Russian of Moscow or of Smolensk) or regional ones, such as Delaware Valley English, Australian English, or Tuscan Italian. Such entities are of unequal rank; South Carolina English, for instance, is included in Southern American English. Regional dialects do have some internal variation, but the differences within a regional dialect are supposedly smaller than differences between two regional dialects of the same rank. In a number of areas ("linguistic landscapes") where the dialectal differentiation is essentially even, it is hardly justified to speak of regional dialects. This uniformity has led many linguists to deny the meaningfulness of such a notion altogether; very frequently, however, bundles of isoglosses—or even a single isogloss of major importance—permit the division of a territory into regional dialects (see Figure 1 for the dialectal division of American English in the Atlantic states). The public is often aware

Adapted from H. Kurath, *Word Geography of the Eastern United States*; the University of Michigan Press

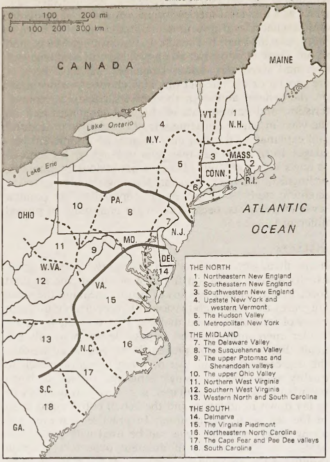


Figure 1: Dialect areas of the eastern United States.

of such divisions, usually associating them with names of geographic regions or provinces, or with some feature of pronunciation; *e.g.*, Southern English or Russian *o*-dialects and *a*-dialects. Especially clear-cut cases of division are those in which geographic isolation has played the principal role; *e.g.*, Australian English or Louisiana French.

Social dialects. Another important axis of differentiation is that of social strata. In many localities, dialectal differences are connected with social classes, educational levels, or both. More highly educated speakers and, often, those belonging to a higher social class tend to use more features belonging to the standard language, whereas the original dialect of the region is better preserved in the speech of the lower and less educated classes. In large urban centres, innovations unknown in the former dialect of the region frequently develop. Thus, in cities the social stratification of dialects is especially relevant and far-reaching, whereas in rural areas, with a conservative way of life, the traditional geographic dialectal differentiation prevails.

Technical terminology and slang

Educational differences among speakers strongly affect the extent of their vocabulary. In addition, practically every profession has its own expressions, which include the technical terminology and sometimes also the casual words or idioms peculiar to the group. Slang, too, is characterized mainly by a specific vocabulary and is much more flexible than an ordinary dialect, as it is subject to fashion and depends strongly on the speaker's age group. Slang—just as a professional dialect—is used mainly by persons who are in a sense bidialectal; *i.e.*, they speak some other dialect or the standard language, in addition to slang. Dialectal differences also often run parallel with the religious or racial division of the population.

Dialectal change and diffusion. The basic cause of dialectal differentiation is linguistic change. Every living language constantly undergoes changes in its various elements. Because languages are extremely complex systems of signs, it is almost inconceivable that linguistic evolution could affect the same elements and even transform them in the same way in all localities where one language is spoken and for all speakers in the same locality. At first glance, differences caused by linguistic change seem to be slight, but they inevitably accumulate with time (*e.g.*, compare Chaucer's English with modern English or Latin with modern Italian, French, Spanish, or Romanian). Related languages usually begin as dialects of the same language.

When a change (an innovation) appears among only one section of the speakers of a language, this automatically creates a dialectal difference. Sometimes an innovation in dialect A contrasts with the unchanged usage (archaism) in dialect B. Sometimes a separate innovation occurs in each of the two dialects. Of course, different innovations will appear in different dialects, so that, in comparison with its contemporaries, no one dialect as a whole can be considered archaic in any absolute sense. A dialect may be characterized as relatively archaic, because it shows fewer innovations than the others; or it may be archaic in one feature only.

After the appearance of a new dialectal feature, interaction between speakers who have adopted this feature and those who have not leads to the expansion or the curtailment of its area or even to its disappearance. In a single social milieu (generally the inhabitants of the same locality, generation, and social class), the chance of the complete adoption or rejection of a new dialectal feature is very great; the intense contact and consciousness of membership within the social group fosters such uniformity. When several age groups or social strata live within the same locality and especially when people speaking the same language live in separate communities, dialectal differences are easily maintained.

Natural barriers and language change

The element of mutual contact plays a large role in the maintenance of speech patterns; that is why differences between geographically distant dialects are normally greater than those between dialects of neighbouring settlements. This also explains why bundles of isoglosses so often form along major natural barriers—impassable mountain ranges, deserts, uninhabited marshes or forests, or wide rivers—or along political borders. Similarly, racial or religious differences contribute to linguistic differentiation because contact between members of one faith or race and those of another within the same area is very often much more superficial and less frequent than contact between members of the same racial or religious group. An especially powerful influence is the relatively infrequent occurrence of intermarriages, thus preventing dialectal mixture at the point where it is most effective; namely, in the mother tongue learned by the child at home.

Unifying influences on dialects. Communication lines such as roads (if they are at least several centuries old), river valleys, or seacoasts often have a unifying influence. Also, important urban centres, such as Paris, Utrecht, or Cologne, often form the hub of a circular region in which approximately the same dialect is spoken. In such areas, the prestige dialect of the city has obviously expanded. As a general rule, those dialects, or at least certain dialectal features, with greater social prestige tend to replace those that are valued lower on the social scale.

In times of less frequent contact between populations, dialectal differences increase; in periods of greater contact, they diminish. The general trend in modern times is for dialectal differences to diminish, above all through the replacement of dialectal traits by those of the standard language. Mass literacy, schools, increased mobility of populations, and, in the last few decades, the evergrowing role of mass communications all contribute to this tendency. Naturally, the extent of such unifying action varies greatly in different linguistic domains, depending on the level of civilization. Nevertheless, the most thorough example of linguistic force exerted by a single dominating civilization belongs to ancient times: in the Hellenistic era, almost all ancient Greek dialects were replaced by the so-called koine, based on the dialect of Athens.

Mass migrations may also contribute to the formation of a more or less uniform dialect over broad geographic areas. Either the resulting dialect is that of the original homeland of a particular migrating population or it is a dialect mixture formed by the levelling of differences among migrants from more than one homeland. The degree of dialectal differentiation depends to a great extent on the length of time a certain population has remained in a certain place. Thus, it is understandable that the diversification of the English language is far greater in the British Isles than, for example, in North America (especially if the number of dialectal differences is considered on a comparable area basis, such as how many per 1,000 square miles). In the U.S. itself much greater diversity is evident among dialects in old colonial America—along the Atlantic coast—than among dialects west of the Appalachians. It is also typical that phonological differences are more far-reaching in Switzerland among Swiss-German dialects than throughout the vast territory where the Russian language is spoken, extending from St. Petersburg to eastern Siberia. Such a situation results not only from migrations of the Russian population, (as compared to the centuries of Swiss stability) but also from the contrasting geographical configurations: in Russia, there is unobstructed communication in all directions; in mountainous Switzerland, the territory is carved into small, isolated units.

Dialect differentiation and population movements

Migrations and, more rarely, geographical phenomena may in some areas cause a much stronger dialectal differentiation in one direction than in others. Isoglosses in the U.S., for example, run predominantly in an east-west direction, reflecting the westward stream of migration during the colonization of areas west of the Appalachians. Similarly, the majority of isoglosses in Russia follow latitude, but in the opposite (west-east) direction.

Focal, relic, and transitional areas. Dialectologists often distinguish between focal areas—which provide sources of numerous important innovations and usually coincide with centres of lively economic or cultural activity—and relic areas—places toward which such innovations are spreading but have not usually arrived. (Relic areas also have their own innovations, which, however, usually extend over a smaller geographical area.) Relic areas or relic phenomena are particularly common in out-of-the-way regional pockets or along the periphery of a particular language's geographical territory. An example of a focal area in the U.S. would be the Boston region, while rural Maine and New Hampshire and Cape Cod and Nantucket Island would be typical relic areas (see Figure 2).

The borders of regional dialects often contain transitional areas that share some features with one neighbour and some with the other. Such mixtures result from unequal diffusion of innovations from both sides. Similar unequal diffusion in mixed dialects in any region also may be a consequence of population mixture created by migrations.

In regions with many bilingual speakers (*e.g.,* along the border between two languages) dialects of both languages will often undergo changes influenced by the other tongue. This is manifested not only in numerous loanwords but often also in the adoption of phonological or grammatical features. Such phenomena are particularly frequent in a population that once spoke one language and only later adopted the second language. In extreme cases, a so-called creolized language develops. (Creoles are pidgin languages that have become the only or major language of a speech community. See LANGUAGES OF THE WORLD: *Pidgin.*)

Written form of dialects

Standard languages. Standard languages arise when a certain dialect begins to be used in written form, normally throughout a broader area than that of the dialect itself. The ways in which this language is used—in administrative matters, literature, economic life—lead to the minimization of linguistic variation. The social prestige attached to the speech of the richest, most powerful, and most highly educated members of a society transforms their language into a model for others; it also contributes to the elimination of deviating linguistic forms. Dictionaries and grammars help to stabilize linguistic norms, as do the activity of scholarly institutions and, sometimes, governmental intervention. The base dialect for a country's standard language is very often the original dialect

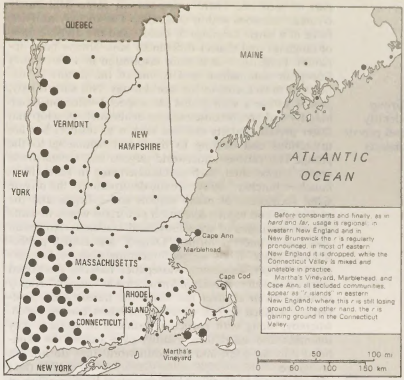


Figure 2: New England pronunciation of preconsonantal and final *r*. The largest circles indicate regular use of this *r*, the smallest ones sporadic use, and the two intermediate sizes rather evenly divided usage.

Adapted from A.J. Bronstein, *The Pronunciation of American English—An Introduction to Phonetics* (1960), Appleton-Century-Crofts; with permission from the American Council of Learned Societies

of its capital—in France, Paris; in England, London; in Russia, Moscow. Or the base may be a strong economic and cultural centre—in Italy, Florence. Or the language may be a combination of several regional dialects—*e.g.,* German or Polish.

Even a standard language that was originally based on one local dialect changes, however, as elements of other dialects infiltrate into it over the years. The actual development in any one linguistic area depends on historical events. Sometimes even the distribution of standard languages may not correspond to the dialectal situation. Dutch and Flemish dialects are a part of the Low German dialectal area, which embraces all of northern Germany, as well as The Netherlands and part of Belgium. In one part of the dialectal area, however, the standard language is based on High German, and, in the other part, the standard language is Dutch or Flemish, depending on the nationality of the respective populations. In the U.S., where there is no clearly dominant political or cultural centre—such as London or Paris—and where the territory is enormous, the so-called standard language shows perceptible regional variations in pronunciation.

In most developed countries, the majority of the population has an active (speaking, writing) or at least passive (understanding) command of the standard language. Very often the rural population, and not uncommonly the lower social strata of the urban population as well, are in reality bidialectal. They speak their maternal dialect at home and with friends and acquaintances in casual contacts, and they use the standard language in more formal situations. Even the educated urban population in some regions uses the so-called colloquial language informally. In the German-, Czech-, and Slovene-speaking areas of middle Europe, for example, a basically regional dialect from which the most striking local features have been eliminated is spoken. The use of this type of language is supported by psychological factors, such as feelings of solidarity with a certain region and pride in its traditions or the relaxed mood connected with informal behaviour. (P.I./Ed.)

SLANG

The above-mentioned social and regional dialect variations within languages are more or less natural; that is, they arise largely from the conditions of language use and language transmission, without deliberate intent on the part of speakers. There are, however, some deliberately created variations within languages. The socially unifying force of a single language or dialect and the divisive force of language and dialect differences have always been apparent. Language was a main inspiration in 19th-century European nationalism and is one of the factors in nationalist movements in the world today. Not surprisingly, groups within a society that set a special value on group identity and group consciousness deliberately develop and foster private dialects that are known to insiders but are mysterious and baffling to those not belonging to the group. The various underworld jargons and special trade argots can be cited, such as Loucherbème in Paris (from *boucher* "butcher," by systematic deformation), the private slang vocabulary of many schools and colleges, and the systematic and regular alteration of certain words in some private languages of schoolboys and students—American "pig Latin" and the former "Oxford" and British public-school "-agger" talk ("nogger," meaning agnostic; "wagger pagger bagger," meaning wastepaper basket), developed from universally recognized forms such as "rugger," Rugby football. From all over the world the transient but zealously cherished speech styles of teenagers can be noted, whereby, along with particular fashions in clothes, they maintain and assert their identity over against the childhood they have left and the adult world they do not yet recognize. (Ed.)

Group identity and private dialects

Development of slang. Slang emanates from conflicts in values, sometimes superficial, often fundamental. When an individual applies language in a new way to express hostility, ridicule, or contempt, often with sharp wit, he may be creating slang, but the new expression will perish unless it is picked up by others. If the speaker is a member of a group that finds that his creation projects the emotional reaction of its members toward an idea, person, or social institution, the expression will gain currency according to the unanimity of attitude within the group. A new slang term is usually widely used in a subculture before it appears in the dominant culture. Thus slang—*e.g.,* "sucker," "honkey," "shave-tail," "jerk"—expresses the attitudes, not always derogatory, of one group or class toward the values of another. Slang sometimes stems from within the group, satirizing or burlesquing its own values, behaviour, and attitudes; *e.g.,* "shotgun wedding," "cake eater," "greasy spoon." Slang, then, is produced largely by social forces rather than by an individual speaker or writer who, single-handed (like Horace Walpole, who coined "serendipity" more than 200 years ago), creates and establishes a word in the language. This is one reason why it is difficult to determine the origin of slang terms.

The specialized languages of subcultures

Creators of slang. Civilized society tends to divide into a dominant culture and various subcultures that flourish within the dominant framework. The subcultures show specialized linguistic phenomena, varying widely in form and content, that depend on the nature of the groups and their relation to each other and to the dominant culture. The shock value of slang stems largely from the verbal transfer of the values of a subculture to diametrically opposed values in the dominant culture. Names such as fuzz, pig, fink, bull, and dick for policemen were not created by officers of the law. (The humorous "dickless tracy," however, meaning a policewoman, *was* coined by male policemen.)

Occupational groups are legion, and while in most respects they identify with the dominant culture, there is just enough social and linguistic hostility to maintain group solidarity. Terms such as scab, strike-breaker, company-man, and goon were highly charged words in the era in which labour began to organize in the United States; they are not used lightly even today, though they have been taken into the standard language.

In addition to occupational and professional groups, there are many other types of subcultures that supply slang. These include sexual deviants, narcotic addicts, ghetto groups, institutional populations, agricultural subsocieties, political organizations, the armed forces, Gypsies, and sports groups of many varieties. Some of the most fruitful sources of slang are the subcultures of professional criminals who have migrated to the New World since the 16th century. Old-time thieves still humorously refer to themselves as FFV—First Families of Virginia.

In criminal subcultures, pressure applied by the dominant culture intensifies the internal forces already at work, and the argot forming there emphasizes the values, attitudes, and techniques of the subculture. Criminal groups seem to evolve about this specialized argot, and both the subculture and its slang expressions proliferate in response to internal and external pressures.

Sources. Most subcultures tend to draw words and phrases from the contiguous language (rather than creating many new words) and to give these established terms new and special meanings; some borrowings from foreign languages, including the American Indian tongues, are traditional. The more learned occupations or professions like medicine, law, psychology, sociology, engineering, and electronics tend to create true neologisms, often based on Greek or Latin roots, but these are not major sources for slang, though nurses and medical students adapt some medical terminology to their slang, and air force personnel and some other branches of the armed services borrow freely from engineering and electronics.

Linguistic processes forming slang. The processes by which words become slang are the same as those by which other words in the language change their form or meaning or both. Some of these are the employment of metaphor, simile, folk etymology, distortion of sounds in words, generalization, specialization, clipping, the use of acronyms, elevation and degeneration, metonymy, synecdoche, hyperbole, borrowings from foreign languages, and the play of euphemism against taboo. The English word trip is an example of a term that has undergone both specialization and generalization. It first became specialized to mean a psychedelic experience resulting from the drug LSD. Subsequently, it generalized again to mean any experience on any drug, and beyond that to any type of "kicks" from anything. Clipping is exemplified by the use of "grass" from "laughing grass," a term for marijuana. "Funky," once a very low term for body odour, has undergone elevation among jazz buffs to signify "the best"; "fanny," on the other hand, once simply a girl's name, is currently a degenerated term that refers to the buttocks (in England, it has further degenerated into a taboo word for the female genitalia). There is also some actual coinage of slang terms.

Characteristics of slang. Psychologically, most good slang harks back to the stage in human culture when animism was a worldwide religion. At that time, it was believed that all objects had two aspects, one external and objective that could be perceived by the senses, the other imperceptible (except to gifted individuals) but identical with what we today would call the "real" object. Human survival depended upon the manipulation of all "real" aspects of life—hunting, reproduction, warfare, weapons, design of habitations, nature of clothing or decoration, etc.—through control or influence upon the *animus,* or imperceptible phase of reality. This influence was exerted through many aspects of sympathetic magic, one of the most potent being the use of language. Words, therefore, had great power, because they evoked the things to which they referred.

Primitive origins of slang

Civilized cultures and their languages retain many remnants of animism, largely on the unconscious level. In Western languages, the metaphor owes its power to echoes of sympathetic magic, and slang utilizes certain attributes of the metaphor to evoke images too close for comfort to "reality." For example, to refer to a woman as a "broad" is automatically to increase her girth in an area in which she may fancy herself as being thin. Her reaction may, thus, be one of anger and resentment, if she happens to live in a society in which slim hips are considered essential to feminine beauty. Slang, then, owes much of its power to shock to the superimposition of images that are incongruous with images (or values) of others, usually members of the dominant culture. Slang is most popular when its imagery develops incongruity bordering on social satire. Every slang word, however, has its own history and reasons for popularity. When conditions change, the term may change in meaning, be adopted into the standard language, or continue to be used as slang within certain enclaves of the population. Nothing is flatter than dead slang. In 1910, for instance, "Oh you kid" and "23-skiddoo" were quite stylish phrases in the U.S. but they have gone with the hobble skirt. Children, however, unaware of anachronisms, often revive old slang under a barrage of older movies rerun on television.

Transience and persistence

Some slang becomes respectable when it loses its edge; "spunk," "fizzle," "spent," "hit the spot," "jazz," "funky," and "p.o.'d," once thought to be too indecent for feminine ears, are now family words. Other slang survives for centuries, like "bones" for dice (Chaucer), "beat it" for run away (Shakespeare), "duds" for clothes, and "booze" for liquor (Dekker). These words must have been uttered as slang long before appearing in print, and they have remained slang ever since. Normally, slang has both a high birth and death rate in the dominant culture, and excessive use tends to dull the lustre of even the most colourful and descriptive words and phrases. The rate of turnover in slang words is undoubtedly encouraged by the mass media, and a term must be increasingly effective to survive.

While many slang words introduce new concepts, some of the most effective slang provides new expressions—fresh, satirical, shocking—for established concepts, often very respectable ones. Sound is sometimes used as a basis for this type of slang, as, for example, in various phonetic distortions (*e.g.*, pig Latin terms). It is also used in rhyming slang, which employs a fortunate combination of both sound and imagery. Thus, gloves are "turtledoves" (the gloved hands suggesting a pair of billing doves), a girl is a "twist and twirl" (the movement suggesting a girl walking), and an insulting imitation of flatus, produced by blowing air between the tip of the protruded tongue and the upper lip, is the "raspberry," cut back from "raspberry tart." Most slang, however, depends upon incongruity of imagery, conveyed by the lively connotations of a novel term applied to an established concept. Slang is not all of equal quality, a considerable body of it reflecting a simple need to find new terms for common ones, such as the hands, feet, head, and other parts of the body. Food, drink, and sex also involve extensive slang vocabulary. Strained or synthetically invented slang lacks verve, as can be seen in the desperate efforts of some sportswriters to avoid mentioning the word baseball—*e.g.*, a batter does not hit a baseball but rather "swats the horsehide," "plasters the pill," "hefts the old apple over the fence," and so on.

The most effective slang operates on a more sophisticated level and often tells something about the thing named, the person using the term, and the social matrix against which it is used. Pungency may increase when full understanding of the term depends on a little inside information or knowledge of a term already in use, often on the slang side itself. For example, the term Vatican roulette (for the rhythm system of birth control) would have little impact if the expression Russian roulette were not already in wide usage.

Diffusion of slang. Slang invades the dominant culture as it seeps out of various subcultures. Some words fall dead or lie dormant in the dominant culture for long periods. Others vividly express an idea already latent in the dominant culture and these are immediately picked up and used. Before the advent of mass media, such terms invaded the dominant culture slowly and were transmitted largely by word of mouth. Thus a term like snafu, its shocking power softened with the explanation "situation normal, all fouled up," worked its way gradually from the military in World War II by word of mouth (because the media largely shunned it) into respectable circles. Today, however, a sportscaster, news reporter, or comedian may introduce a lively new word already used by an in-group into millions of homes simultaneously, giving it almost instant currency. For example, the term uptight was first used largely by criminal narcotic addicts to indicate the onset of withdrawal distress when drugs are denied. Later, because of intense journalistic interest in the drug scene, it became widely used in the dominant culture to mean anxiety or tension unrelated to drug use. It kept its form but changed its meaning slightly.

Slang and the mass media

Other terms may change their form or both form and meaning, like "one for the book" (anything unusual or unbelievable). Sportswriters in the U.S. borrowed this term around 1920 from the occupational language of then legal bookmakers, who lined up at racetracks in the morning ("the morning line" is still figuratively used on every sports page) to take bets on the afternoon races. Newly arrived bookmakers went to the end of the line, and any bettor requesting unusually long odds was motioned down the line with the phrase, "That's one for the end book." The general public dropped the "end" as meaningless, but old-time gamblers still retain it. Slang spreads through many other channels, such as popular songs, which, for the initiate, are often rich in double entendre.

When subcultures are structurally tight, little of their language leaks out. Thus the Mafia, in more than a half-century of powerful criminal activity in America, has contributed little slang. When subcultures weaken, contacts with the dominant culture multiply, diffusion occurs, and their language appears widely as slang. Criminal narcotic addicts, for example, had a tight subculture and a highly secret argot in the 1940s; now their terms are used freely by middle-class teenagers, even those with no real knowledge of drugs.

Slang expressing emotional attitudes

Uses of slang. Slang is used for many purposes, but generally it expresses a certain emotional attitude; the same term may express diametrically opposed attitudes when used by different people. Many slang terms are primarily derogatory, though they may also be ambivalent when used in intimacy or affection. Some crystallize or bolster the self-image or promote identification with a class or in-group. Others flatter objects, institutions, or persons but may be used by different people for the opposite effect. "Jesus freak," originally used as ridicule, was adopted as a title by certain street evangelists. Slang sometimes insults or shocks when used directly; some terms euphemize a sensitive concept, though obvious or excessive euphemism may break the taboo more effectively than a less decorous term. Some slang words are essential because there are no words in the standard language expressing exactly the same meaning; *e.g.*, "freak-out," "barn-storm," "rubberneck," and the noun "creep." At the other extreme, multitudes of words, vague in meaning, are used simply as fads.

There are many other uses to which slang is put, according to the individual and his place in society. Since most slang is used on the spoken level, by persons who probably are unaware that it is slang, the choice of terms naturally follows a multiplicity of unconscious thought patterns. When used by writers, slang is much more consciously and carefully chosen to achieve a specific effect. Writers, however, seldom invent slang.

It has been claimed that slang is created by ingenious individuals to freshen the language, to vitalize it, to make the language more pungent and picturesque, to increase the store of terse and striking words, or to provide a vocabulary for new shades of meaning. Most of the originators and purveyors of slang, however, are probably not conscious of these noble purposes and do not seem overly concerned about what happens to their language.

Attitudes toward slang. With the rise of naturalistic writing demanding realism, slang began to creep into English literature even though the schools waged warfare against it, the pulpit thundered against it, and many women who aspired to gentility and refinement banished it from the home. It flourished underground, however, in such male sanctuaries as lodges, poolrooms, barbershops, and saloons.

Slang in literature

By 1925 a whole new generation of U.S. and European naturalistic writers was in revolt against the Victorian restraints that had caused even Mark Twain to complain, and today any writer may use slang freely, especially in fiction and drama. It has become an indispensable tool in the hands of master satirists, humorists, and journalists. Slang is now socially acceptable, not just because it is slang but because, when used with skill and discrimination, it adds a new and exciting dimension to language. At the same time, it is being seriously studied by linguists and other social scientists as a revealing index to the culture that produces and uses it. (D.W.M./Ed.)

OTHER SPECIALIZED LANGUAGES

Jargon. Sometimes, as in the case of criminal argots, part of the function of special languages is deliberately to mislead and obstruct the rest of society and the authorities in particular; they may even become wholly impenetrable to outsiders. But this is not the sole or main purpose of most specialized varieties of language. Professions whose members value their standing in society and are eager to render their services to the public foster their own vocabulary and usage, partly to enhance the dignity of their profession and the skills they represent but partly also to increase their efficiency. An example of this is the language of the law and of lawyers.

The cultivation and maintenance of specialized types of language by certain professions should not be regarded as trivially or superficially motivated. In general usage, languages are necessarily imprecise, or they would lack the flexibility and infinite extensibility demanded of them. But for certain purposes in restricted situations much greater precision is required, and part of the function of the particular style and vocabulary of legal language is the avoidance, so far as may be possible, of all ambiguity and the explicit statement of all necessary distinctions. This is why legal texts, when read out of their context, seem so absurdly pedantic and are an easy target for ridicule. Similar provision for detail and clarity characterizes the specialist jargons of medicine and of the sciences in general and also of philosophy. Indeed, one might regard the formulas of modern symbolic logic as the result of a consciously developed and specialized written language for making precise the relations of implication and inference between statements that, when couched in everyday language, are inexact and open to misinterpretation. Some would go as far as to say that traditional metaphysics is no more than the result of misunderstanding everyday discourse and that the main purpose of philosophy is to resolve the puzzles that arise from such misunderstandings.

Stereotyped language of games

The use of specialized types of language in fostering unity is also evidenced in the stereotyped forms of vocabulary employed in the playing of certain games. Tennis scores use the sequence "love, 15, 30, 40, and game"; cricketers verbally appeal to the umpire when a batsman may be out by calling "How's that?" and the ways of being out are designated by stereotypes, "run out," "leg before wicket," "stumped," and so forth. The esoteric language of horse racing and its associated wagering of money is well known, though not readily understood by outsiders.

The ancient but persistent recognition of the power of language is apparent in the respect for correctness in the use of language in any sphere of life having supernatural connections. Those credited with such connections employ special formulas and rigidly prescribed modes of diction; examples of the language of magic and of magicians are widespread, ranging from the usages of shamans and witch doctors to the ritual "abracadabra" of the mock magic displayed by conjurors at children's parties.

The efficacy of religious worship and of prayers is frequently associated with the strict maintenance of correct forms of language, taught by priests to their successors, lest the ritual become invalid. In ancient India the preservation in all its supposed purity of the language used in the performance of certain religious rituals (Sanskrit) gave rise to one of the world's most important schools of linguistics and phonetics. In the Christian churches one can observe the value placed by Church of England and Episcopalian churchmen on the formal English of the Authorized Version of the Bible and of *The Book of Common Prayer*, despite recent attempts at replacing these ritual forms of language by forms taken from modern spoken vernaculars.

Pidgins and creoles. Some specialized languages were developed to keep the outsider at bay. In other circumstances, languages have been deliberately created to facilitate communication with outsiders. This happens when people speaking two different languages have to work together, usually in some form of trade relation or administrative routine. In such situations the so-called pidgins arise, more or less purposively made up of vocabulary items from each language, with mutual abandonment of grammatical complexities that would cause confusion to either party. Pidgins have been particularly associated with areas settled by European traders; examples have been Chinook Jargon, a lingua franca based on an American Indian language and English and formerly used in Washington and Oregon, and Beach-la-mar, an English-based pidgin of parts of the South Seas.

Creolization of a pidgin language

Sometimes, as the result of relatively permanent settlement and the intermixture of two speech communities, a pidgin becomes the first language, or mother tongue, of later generations, ultimately displacing both the original languages. First languages arising in this way from artificially created pidgins are called creoles. Notable among creoles is the language of Haiti, Haitian Creole, built up from the French of the settlers and the African language of the former slaves; it shows lexical and grammatical features of both sources.

Creoles differ from pidgins in that, as first languages, they are subject to the natural processes of change like any other language (see below *Linguistic change*); and, despite the deliberately simplified form of the original pidgin, in the course of generations creoles develop their own complexities. The reason is plain to see. The restricted uses to which pidgins were first put and for which they were devised did not require any great flexibility. Once such a language becomes the first or only language of many people, it must perforce acquire the resources (*i.e.*, the complexity) to respond adequately to all the requirements of a natural language (see also LANGUAGES OF THE WORLD).

Nonverbal language. Speech and writing are, indeed, the fundamental faculties and activities referred to by the term language. There are, however, areas of human behaviour for which the term is used in a peripheral and derivative sense.

When individuals speak, they do not normally confine themselves to the mere emission of speech sounds. Because speaking usually involves at least two parties in sight of each other, a great deal of meaning is conveyed by facial expression, tone of voice, and movements and postures of the whole body but especially of the hands; these are collectively known as gestures. The contribution of bodily gestures to the total meaning of a conversation is in part culturally determined and differs in different communities. Just how important these visual symbols are may be seen when one considers how much less effective telephone conversation is as compared with conversation face to face; the experience of involuntarily smiling at the telephone receiver and immediately realizing that this will convey nothing to the hearer at the other end of the line is common. Again, the part played in emotional contact and in the expression of feelings by facial expressions and tone of voice, quite independently of the words used, has been shown in tests in which subjects have been asked to react to sentences that appear as friendly and inviting when read but are spoken angrily and, conversely, to sentences that appear as hostile but are spoken with friendly facial expressions. It is found that it is the visual accompaniments and tone of voice that elicit the main emotional response. A good deal of what goes under the heading of sarcasm exploits these contrasts.

Expression by vocal, non-linguistic utterances

Just as there are paralinguistic activities such as facial expressions and bodily gestures integrated with and assisting the communicative function of spoken language, so there are vocally produced noises that cannot be regarded as part of any language, though they help in communication and in the expression of feeling. These include laughter, shouts and screams of joy, fear, pain, and so forth, and conventional expressions of disgust, triumph, and so on, traditionally spelled "ugh!," "ha ha!," etc., in English. Such nonlexical ejaculations differ in important respects from language: they are much more similar in form and meaning throughout mankind as a whole, in contrast to the great diversity of languages; they are far less arbitrary than most of the lexical components of language; and they are much nearer the cries of animals produced under similar circumstances and, as far as is known, serve similar expressive and communicative purposes. As noted above, some people have tried to trace the origin of language itself to them.

A language is a symbol system. It may be regarded, because of its infinite flexibility and productivity, as the symbol system *par excellence*. But there are other symbol systems recognized and institutionalized in the different cultures of mankind. Examples of these exist on maps and blueprints and in the conventions of representational art (*e.g.*, the golden halos around the heads of saints in religious paintings). Other symbol systems are musical notation and dance notation, wherein graphic symbols designate musical pitches and other features of musical performance and the movements of formalized dances. More loosely, because music itself can convey and arouse emotions and certain musical forms and structures are often associated with certain types of feeling, one frequently reads of the "language of music" or even of "the grammar of music." The terms language and grammar are here being used metaphorically, however, if only because no symbol system other than language has the same potential of infinite productivity, extension, and precision.

Languages are used by human beings to talk and write to other human beings. Derivatively, bits of languages may be used by humans to control machinery, as when different buttons and switches are marked with words or phrases designating their functions. A recent and specialized development of man-machine language is seen in the various "computer languages" (Cobol, Algol, and Fortran, for example) now in use. These are referred to as programming languages, and they provide the means whereby sets of "instructions" and data of various kinds can be supplied to computers in forms acceptable to these machines. Various types of such languages are employed for different purposes. The development and use of computer languages must now be regarded as a distinct science in itself (for more information, see COMPUTERS: *Digital computer programming* and COMPUTER SCIENCE: *Programming languages and methodology*).

Physiological and physical basis of speech

Primacy of speech in language

For an adequate understanding of human language, it is necessary to keep in mind the absolute primacy of speech. In societies in which literacy is all but universal and language teaching at school begins with reading and writing in the mother tongue, one is apt to think of language as a writing system that may be pronounced. In point of fact, language is a system of spoken communication that may be represented in various ways in writing.

The human being has almost certainly been in some sense a speaking animal from early in the emergence of *Homo sapiens* as a recognizably distinct species. The earliest known systems of writing go back perhaps some 5,000 years. This means that for many hundreds of thousands of years human languages were transmitted from generation to generation and were developed entirely as spoken means of communication. Moreover, in the world as it is today, literacy is still the privilege of a minority in many language communities. Even when literacy is widespread, some languages remain unwritten if they are not economically or culturally important enough to justify creating an alphabet for them and teaching them; then literacy is acquired in a second language learned at school. Such is the case with many speakers of South American Indian languages, who become literate in Spanish or Portuguese. A similar situation prevails in some parts of Africa, where reading and writing are taught in languages spoken over relatively wide areas. In all communities, speaking is learned by children before writing, and all people act as speakers and hearers much more than as writers and readers.

It is, moreover, a total fallacy to suppose that the languages of illiterate or so-called primitive peoples are less structured, less rich in vocabulary, and less efficient than the languages of literate civilizations. The lexical content of languages varies, of course, according to the culture and the needs of their speakers, but observation bears out the statement that the U.S. anthropological linguist Edward Sapir made in 1921: "When it comes to linguistic form, Plato walks with the Macedonian swineherd, Confucius with the head-hunting savage of Assam."

All this means that the structure and composition of language and of all languages have been conditioned by the requirements of speech, not those of writing. Languages are what they are by virtue of their spoken, not their written, manifestations. The study of language must be based on a knowledge of the physiological and physical nature of speaking and hearing. The details of these aspects of language are covered in SPEECH: *The phonetics of speech* and *The physiology of speech;* only the essentials are given here.

SPEECH PRODUCTION

Speaking is in essence the by-product of a necessary bodily process, the expulsion from the lungs of air charged with carbon dioxide after it has fulfilled its function in respiration. Most of the time one breathes out silently; but it is possible, by adopting various postures and by making various movements within the vocal tract, to interfere with the egressive airstream so as to generate noises of different sorts. This is what speech is made of.

Vocal tract and speech organs

The vocal tract comprises the passage from the trachea (windpipe) to the orifices of the mouth and nose; all the organs used in speaking lie in this passage. Conventionally, these are called the organs of speech, and the use in several languages of the same word for the tongue as a part of the body and for language shows the awareness people have of the role played by this part of the mouth in speaking. But few if any of the major organs of speech are exclusively or even mainly concerned with speaking. The lips, the tongue, and the teeth all have essential functions in the bodily economy, quite apart from talking; to think, for example, of the tongue as an organ of speech in the same way that the stomach is regarded as the organ of digestion is fallacious. Speaking is a function superimposed on these organs, and the material of speech is a waste product, spent air, exploited to produce perhaps the most wonderful by-product ever created.

Relatively few types of speech sounds are produced by other sources of air movement; the clicks in some South African languages are examples, and so is the fringe linguistic sound used in English to express disapproval, conventionally spelled "tut." In all languages, however, the great majority of speech sounds have their origin in air expelled through the contraction of the lungs. Air forced through a narrow passage or momentarily blocked and then released creates noise, and characteristic components of speech sounds are types of noise produced by blockage or narrowing of the passage at different places.

If the vocal cords (really more like two curtains) are held taut as the air passes through them, the resultant regular vibrations in the larynx produce what is technically called voice, or voicing. These vibrations can be readily observed by contrasting the sounds of *f* and *v* or of *s* and *z* as usually pronounced; "five" and "size" each begin and end with voiceless and voiced sounds, respectively, which are otherwise formed alike, with the tongue and the lips in the same position. Most consonant sounds and all vowel sounds in English and in the majority of languages are voiced, and voice, in this sense, is the basis of singing and of the rise and fall in speaking that is called intonation, as well as of the tone distinctions in tone languages. The vocal cords may be drawn together more or less tightly, and the vibrations will be correspondingly more or less frequent. A rise in frequency causes a rise in perceived vocal pitch. Speech in which voice is completely excluded is called whispering.

Above the larynx, places of articulation in frequent use are between the back of the tongue and the soft palate, between the blade of the tongue and the ridge just behind the upper front teeth, and between the lips. Stoppage and release (technically, plosion) at these places form the *k* (often written as *c*, "cat"), *t*, and *p* sounds in English and, when voicing is also present, the *g* (as in "gay"), *d*, and *b* sounds. Obstruction at these and other places sufficient to cause noise gives rise to what are called fricative sounds; in English these include the normal pronunciations of *s*, *z*, *f*, and *v* and the *th* sounds in "thin" and "then." A vowel is characterized as the product of the shape of the entire tract between the lips and larynx, without local obstruction though usually with voicing from the vocal

Formation of vowels

cords. It is contrasted with a consonant, though the exact division between these two categories of speech sound is not always easy to draw. Different shaping of the tract produces the different vowel sounds of languages.

The soft palate may be raised or lowered. It is lowered in breathing and allows air to pass in and out through the nose. In the utterance of most speech sounds it is raised, so that air passing through the mouth alone forms the sound; if it is lowered, air passes additionally or alternatively through the nose, producing nasal sounds. All but a few languages have nasal consonants (the English sounds *m*, *n*, and *ng* as in "sing"), and some, such as French, have nasalized vowels as well. A few people regularly allow air to pass through their nasal passages while they speak; such persons are said to "speak through the nose."

All articulatory movements, including the initial expulsion of air from the lungs, may be made with greater or less vigour, giving rise to louder or softer speech or to greater loudness on one part of what is said.

Every different configuration and movement of the vocal tract creates corresponding differences in the air vibrations that comprise and transmit sound. These vibrations, like those of all noises, extend outward in all directions from the source, gradually decreasing to zero or to below the threshold of audibility. They are called sound waves, and they consist of rapid rises and falls in air pressure. The speed at which pressure rises and falls is the frequency. Speech sounds involve complex waves containing vibrations at a number of different frequencies, the lowest being the voice pitch of singing and intonation, produced by the vocal cords in voiced sounds.

The eardrum responds to the different frequencies of speech, provided they retain enough energy, or amplitude (*i.e.*, are still audible). The different speech sounds that make up the utterances of any language are the result of the different impacts on one's ears made by the different complexes of frequencies in the waves produced by different articulatory processes. As the result of careful and detailed observation of the movements of the vocal organs in speaking, aided by various instruments to supplement the naked eye, a great deal is now known about the processes of articulation. Other instruments have provided much information about the nature of the sound waves produced by articulation. Speech sounds have been described and classified both from an articulatory viewpoint, in terms of how they are produced, and from an acoustic viewpoint, by reference to the resulting sound waves (their frequencies, amplitudes, and so forth). Articulatory descriptions are more readily understood, being couched in terms such as nasal, bilabial lip-rounded, and so on. Acoustic terminology requires a knowledge of the technicalities involved for its comprehension. In that almost every person is a speaker and a hearer, it is clear that both sorts of description and classification are important, and each has its particular value for certain parts of the scientific study of language.

Articulatory and acoustic descriptions of speech

LANGUAGE ACQUISITION

In regard to the production of speech sounds, all humans are physiologically alike. It has been shown repeatedly that children learn the language of those who bring them up from infancy. In most cases these are the biological parents, especially the mother, but one's first language is acquired from environment and learning, not from physiological inheritance. Adopted infants, whatever their race or physical type and whatever the language of their actual parents, acquire the language of the adoptive parents.

Different shapes of lips, throat, and other parts of the vocal tract have an effect on voice quality; this is part of the individuality of each person's voice referred to above. Physiological differences, including size of throat and larynx, both overall and in relation to the rest of the vocal tract, are largely responsible for the different pitch ranges characteristic of men's, women's, and children's speech. These differences do not affect one's ability or aptitude to speak any particular language.

Speech is species-specific to humankind. Physiologically, animal communications systems are of all sorts. The animal sounds superficially most resembling speech, the imitative cries of parrots and some other birds, are produced by very different physiological means: birds have no teeth or lips but vocalize by means of the syrinx, a modification of the windpipe above the lungs. Almost all mammals and many other animal species make vocal noises and evince feelings thereby and keep in contact with each other through a rudimentary sort of communication, but those members of the animal kingdom nearest to humans genetically, the great apes, lack the anatomic apparatus necessary for speech.

The development of speech has been linked to upright posture and the freeing of the vocal cords from the frequent need to "hold one's breath" in using the arms for locomotion. Certainly, speaking and hearing—as a primary means of communication—have a number of striking advantages: speech does not depend on daylight or on mutual visibility, it can operate in all directions over reasonably wide areas, and it can be adjusted in loudness to cope with distance. As is seen in crowded rooms, it is possible to pick out some one person's voice despite a good deal of other noise and in the midst of other voices speaking the same language. Also, the physical energy required in speaking is extremely small in relation to the immense power wielded by speech in human life, and scarcely any other activity, such as running, walking, or tool using, interferes seriously with the process.

Advantages of speech as a means of communication

The characteristics just outlined pertain to all of the world's languages. What is more a matter of controversy is the extent to which biological inheritance is involved in language acquisition and language use. The fact that language traditionally has been viewed as species-specific to human beings argues an essential cerebral or mental component, and in the 19th century certain aspects of speech control and use were located in a particular part of the human brain (Broca's convolution).

Whether or not the great apes have the mental capacity to acquire at least a rudimentary form of language has developed into an area of active research. While apes lack the anatomic structures that are necessary for the vocalization of human speech, many investigators nevertheless claim to have taught chimpanzees, gorillas, and orangutans to communicate in languages whose "words" are composed of hand signs or geometric symbols. These claims have been hotly disputed, with critics arguing that the apes have not demonstrated true language acquisition in the sense of understanding the "words" as symbolic abstractions that can be used in new and grammatically meaningful constructions. Researchers working with the apes, however, maintain that at least some of the apes have learned to understand and manipulate the "words" as abstractions.

No one inherits the ability to speak a particular language, but normal children are born with the ability and the drive to acquire a language—namely, the one to which they are predominantly exposed from infancy. Children bring to this task considerable innate ability, because their exposure is largely to a random selection of utterances (apart from any attempts at systematic teaching that they may encounter) occuring within earshot or addressed to them. Yet by late childhood they have, through progressive stages, acquired the basic vocabulary of the language, together with its phonological and grammatical structure. This is substantially the same situation the world over, among literate and illiterate communities, and much the same number of years of childhood is taken up by the process. Thus, it would appear that all languages are roughly equal in complexity and in difficulty of mastery.

It is, therefore, clear that all normal humans bring into the world an innate faculty for language acquisition, language use, and grammar construction. The last phrase refers to the internalization of the rules of the grammar of one's first language from a more or less random exposure to utterances in it. Human children are very soon able to construct new, grammatically acceptable sentences from material they have already heard; unlike the parrot in human society, they are not limited to mere repetition of utterances.

Innate human capacity for language acquisition

What is currently under debate is the part played by this innate ability and its exact nature. Until the 1950s scholars considered language acquisition to be carried out largely by analogical creation from observed patterns of sentences occurring in utterances heard and understood by the child. Such a view, much favoured by persons inclined to a behaviourist interpretation of human learning processes (*e.g.*, the U.S. linguist Leonard Bloomfield), stressed the very evident differences between the structures of different languages, particularly on the surface. Since the late 1950s, a number of linguists have been placing much more emphasis on the inherent grammar-building disposition and competence of the human brain, which is activated by exposure to utterances in a language, especially during childhood, in such a way that it fits the utterances into predetermined general categories and structures. Such linguists, inheritors of the 17th- and 18th-century interest in "universal grammar," put their stress on the underlying similarities of all languages, more especially in the deeper areas of grammatical analysis (for the distinction between deep structure and surface structure in grammar, see the article LINGUISTICS).

Meaning and style in language

The whole object and purpose of language is to be meaningful. Languages have developed and are constituted in their present forms in order to meet the needs of communication in all its aspects.

It is because the needs of human communication are so various and so multifarious that the study of meaning is probably the most difficult and baffling part of the serious study of language. Traditionally, language has been defined, as in the definition quoted above, as the expression of thought, but, as was seen, this involves far too narrow an interpretation of language or far too wide a view of thought to be serviceable. The expression of thought is just one among the many functions performed by language in certain contexts.

TYPES OF MEANING

Structural, or grammatical, meaning. First, one must recognize that the meaning of any sentence comprises two parts, the meanings of the words it contains and the structural or grammatical meaning carried by the sentence itself. In English "the dog chased the cat" and "the boy chased the cat" differ in meaning because "dog" and "boy" are different words with different word meanings; the same applies to equivalent sentences in other languages. The two sentences "the dog chased the cat" and "the cat chased the dog," though containing exactly the same words, are different in meaning because the different word orders distinguish what are conventionally called subject and object. In Latin the two corresponding sentences would be distinguished not by word order, which is grammatically indifferent and largely a matter of style, but by different shapes in the lexical equivalents of "dog" and "cat." In Japanese the grammatical distinction of subject and object, normally marked by the word order subject-object-verb, can be reinforced by a subject particle after the first word and an object particle after the second.

Distinguishing meaning by word order or grammatical form

The formal resources of any language for making distinctions in the structural meanings of sentences are limited by two things: the linear (time) dimension of speaking and the limited memory span of the human brain. Writing copies the time stream of speech with the linear flow of scripts. Diagrams and pictures employ two dimensions, and models employ three; but writing is partially relieved of memory-span restrictions by the permanence of visual marks. Because written texts are almost entirely divorced from oral pronunciation, sentence length and sentence complexity can be carried to extremes, as may be observed in some legal and legislative documents that are virtually unintelligible if read aloud.

Within these linear restrictions, distinctions corresponding to the main uses of language can be made. All languages can employ different sentence structures to state facts (declarative), to ask questions (interrogative), and to enjoin or forbid some course of action (imperative). More delicate means exist to soften or modify these basic distinctions: *e.g.*, "It's cold today, isn't it?"; "Isn't it still raining?"; "Shut the door, would you mind"; "Don't be long, will you?" Languages use their resources differently for these purposes, but, generally speaking, each seems to be equally flexible structurally. The principal resources are word order, word form, syntactic structure, and, in speech, pitch and stress placement. In English, as an example, a word or phrase can be highlighted by being placed first in the sentence when it would not normally occur there: compare "he can't bear loud noises" with "loud noises he can't bear" or "loud noises, he can't bear them." The object noun or noun phrase can also be put first by making the sentence passive; this allows the original subject to be omitted if one does not know or does not want to refer to an agent: "the town was destroyed (by the revolutionaries)." Within and together with all these possibilities, almost any word can be made contrastively prominent by being stressed (spoken more loudly) or by being uttered on a higher pitch, and very often these two are combined: "I asked you for *red* roses (not yellow)"; "I meant it for *you* (not her)"; "*I* know nothing about it (someone else may)." Prominence is especially associated with intonation, itself an important carrier of structural meaning in speech. One may state facts, ask questions, and give instructions with a variety of intonations indicating, along with visible gestures, different attitudes, feelings, and social and personal relations between speaker and hearer.

The possibilities of expressing structural meanings are a most important part of any language. They are acquired along with the rest of one's first language in childhood and are learned more slowly and with more difficulty in mastering a second or later language. Scholars are still only at the beginning of a full formal analysis of these resources, as far as most languages are concerned, and are still further from an adequate understanding of all the semantic functions performed by means of these resources.

Lexical meaning. The other component of sentence meaning is word meaning, the individual meanings of the words in a sentence, as lexical items. The concept of word meaning is a familiar one. Dictionaries list words and in one way or another state their meanings. It is regarded as a sensible question to ask of any word in a language, "What does it mean?" This question, like many others about language, is easier to ask than to answer.

Meanings of words

It is through lexical resources that languages maintain the flexibility their open-ended commitments demand. Every language has a vocabulary of many thousands of words, though not all are in active use, and some are known only to relatively few speakers. Perhaps the commonest delusion in considering vocabularies is the assumption that the words of different languages, or at least their nouns, verbs, and adjectives, label the same inventory of things, processes, and qualities in the world but unfortunately label them with different labels from language to language. If this were so, translation would be easier than it is; but the fact that translation, though often difficult, is possible indicates that people are talking about similar worlds of experience in their various languages.

Languages in part create the world in which men live. Of course, many words do name existing bits and pieces of earth and heaven: "stone," "tree," "dog," "woman," "star," "cloud," and so on. Others, however, do not so much pick out what is there as classify it and organize one's relations with it and with each other with regard to it. A range of living creatures are mammals or are vertebrates, because people classify them in these ways, among others, by applying selected criteria and so determining the denotation of the words mammal and vertebrate. Plants are vegetables or weeds according as groups of people classify them, and different plants are included and excluded by such classifications in different languages and different cultures.

Time and its associated vocabulary ("year," "month," "day," "hour," "minute," "yesterday," "tomorrow," and so on) do not refer to discrete sections of reality but enable people to impose some sort of order, in agreement with others, on the processes of change observed in the world. Personal pronouns pick out the persons speaking, spoken to, and spoken about; but some languages make different distinctions in their pronouns from those made in English. For example, in Malay, *kita*, which means "we," including the person addressed, is distinct from *kami,* a form for "we" that includes the speaker and a third person or persons but excludes the person addressed. In Japanese and in several other languages, a variety of words denoting the 1st and 2nd persons indicate additionally the observed or intended social relationship of those involved.

Other word meanings are even more language and culture bound, and in consequence harder to translate. "Right" and "wrong," "theft," "inheritance," "property," "debt," "sin," and "crime" (as different sorts of wrongdoing) are just a few of the words regulating one's conduct and relations with one's fellows in a particular culture. Translation becomes progressively harder as one moves to languages of more remote cultures, and it has been said that it requires "a unification of cultural context." Insofar as a person's understanding of the universe and of the relations between himself and other people is closely linked with the language he speaks, it must be assumed, and the evidence confirms this assumption, that the child progressively acquires such understanding along with his language.

Onomatopoeic words

The great majority of word shapes bear no direct relation to their lexical meanings. If they did, languages would be more alike. What are called onomatopoeic words are rather similar in shape through different languages: French *coucou,* English "cuckoo," and German *Kuckuck* directly mimic the call of the bird. English "dingdong" and German *bim-bam* share several sound features in common that partially resemble the clanging of bells. More abstractly, some direct "sound symbolism" has been seen between certain sound types and visual or tactile shapes. Most people agree that the made-up word "oomboolu" would better designate a round, bulbous object than a spiky one. In addition, the appropriateness of the vowel sound represented by *ee* in English "wee" and *i* in French *petit* "small" and Italian *piccolo* "small" for expressing things of small size has been traced in several languages.

All this, however, is a very small part of the vocabulary of any language. For by far the largest number of words in a language there is no direct association between sound and meaning. English "horse," German *Pferd,* French *cheval,* Latin *equus,* and Greek *hippos* are all unrelated to the animal so named, except that these words are so used in the languages concerned. This is what is meant by the term arbitrary in the second definition of language quoted at the beginning of this article. Vocabulary has to be largely arbitrary, because the greater part of the world and of man's experience is not directly associated with any kind of noise, and it is a contingent, though universal, fact of history and biology that sound and not the material of some other sense is the basis of human language.

The relations between sentence structure and structural meanings are also largely arbitrary and tacitly conventional. Though loudness and stress for emphasis and certain linguistic indications of anger, excitement, and the like are more closely akin to nonlinguistic ejaculations and are somewhat similar across language divisions, actual intonations and features such as word order, word inflection, and grammatical particles, used in maintaining distinctions in structural meaning, differ markedly in different languages.

SEMANTIC FLEXIBILITY

Not only are word meanings somewhat different in different languages; they are not fixed for all time in any one language. Semantic changes take place all along (see below), and at any moment the semantic area covered by a word is indeterminately bordered and differs from context to context. This is a further aspect and condition of the inherent and necessary flexibility of language.

Precision and imprecision of words

General and specific designations. A person can be as precise or as imprecise as he needs or wishes to be. In general, words are fairly imprecise; yet for particular purposes their meanings can be tightened up, usually by bringing in more words or phrases to divide up a given field in more detail. "Good" contrasts generally with "bad"; but one can, for example, grade students as "first-class," "excellent," "very good," "good," "fair," "poor," and "failed" (or "bad"). In this case, "good" now covers a restricted and relatively low place in a field of associated terms. Colour words get their meanings from their mutual contrasts. The field of visually discriminable hues is very large and goes far beyond the resources of any vocabulary as it is normally used. Children learn the central or basic colour words of their language fairly early and at the same time; such terms as red and green are normally learned before subdivisions such as crimson and scarlet or chartreuse. It is well known that languages make their primary divisions of the spectrum of colours in different places; Japanese *aoi* covers many of the hues referred to in English by "green" and "blue," while "blue" covers much of the range of the two Russian words *goluboy* and *siny.* While the actual colour vocabularies of languages differ, however, recent research by Brent Berlin and Paul Kay has tried to show that "there exist universally for humans eleven basic perceptual color categories" that serve as reference points for the colour words of a language, whatever number may be regularly employed at any time.

Ordinarily, considerable areas of indeterminate designation in colour vocabulary and in other fields are tolerated; between "red" and "purple" and between "purple" and "blue" there are hues that one would hesitate to assign firmly to one or the other and on which there would be considerable personal disagreement. When greater precision than normal is required—as, for example, in listing paint or textile colours—all kinds of additional terms can be brought into service to supplement the usual vocabulary: "off-white," "light cream," "lemon," "blush pink," and so on.

The vocabulary of kinship terms varies from language to language, reflecting cultural differences. English distinguishes the nearer kinsfolk by sex: "mother, father"; "sister, brother"; "aunt, uncle"; and others. Other languages, such as Malay, make a lexical distinction of age the primary one, with separate words for elder brother or sister and younger brother or sister. Still other languages—for example, some American Indian ones—use different words for the sister of a man and for the sister of a woman. But beyond this any language can be as precise as the situation demands in kin designation. When it is necessary, English speakers can specify "elder sister" and "female cousin," and within the overall category it is possible to distinguish "first and second cousins" and "cousins once removed," distinctions that it is ordinarily pedantic to make.

Vocabulary of mathematics

The best example of infinite precision available from a strictly limited lexical stock is in the field of arithmetic. Between any two whole numbers a further fractional or decimal number may always be inserted, and this may go on indefinitely: between 10 and 11, $10\frac{1}{2}$ (10.5), $10\frac{1}{4}$ (10.25), $10\frac{1}{8}$ (10.125), and so on. Thus, the mathematician or the physical scientist is able to achieve any desired degree of quantitative precision appropriate to his purposes; hence the importance of quantitative statements in the sciences—any thermometric scale contains far more distinctions of temperature than are reasonably available in the vocabulary of a language ("hot," "warm," "cool," "tepid," "cold," and so on). For this reason mathematics has been described as the ideal use of language, but for many purposes in everyday life the very imprecision of natural languages is the source of their strength and adaptability.

Neologisms. Every living language can readily be adapted to meet changes occurring in the life and culture of its speakers, and the main weight of such changes falls on vocabulary. Grammatical and phonological structures are relatively stable and change noticeably over centuries rather than decades (see below); but vocabularies can change very quickly both in word stock and in word meanings. Consider as an example the changes wrought by modern technology in the vocabularies of all European languages since 1945. Before that date "transistor" and "cosmonaut" did not exist, and "nuclear disarmament" would scarcely have had any clear meaning.

Every language can alter its vocabulary very easily, which means that every speaker can without effort adopt new words, accept or invent new meanings for existing words, and of course, cease to use some words or cease to use them in certain meanings. Dictionaries list some words and some meanings as "obsolete" or "obsolescent" to indicate this process. No two speakers share precisely the same vocabulary of words readily used and readily understood, though they may speak the same dialect. They will, however, naturally have the great majority of words in their vocabularies in common.

Languages have various resources for effecting changes in vocabulary. Meanings of existing words may change. With the virtual disappearance of falconry as a sport in England, "lure" has lost its original meaning of a bunch of feathers on a string by which hawks were recalled to their handler and is used now mainly in its metaphorical sense of enticement. The additional meaning of "nuclear" has already been mentioned; one may list it with words such as computer and jet, which acquired new ranges of meaning in the mid-20th century.

Creation of new words

All languages have the means of creating new words to bear new meanings. These can be new creations; "Kodak" is one such, invented at the end of the 19th century by George Eastman; "chortle," now in general use, was a jocular creation of the English writer and mathematician Lewis Carroll (creator of *Alice in Wonderland*); and "gas" was formed in the 17th century by the Belgian chemist and physician Jan Baptist van Helmont as a technical term in chemistry, loosely modelled on the Greek *chaos* ("formless void"). But mostly languages follow definite patterns in their innovations. Words can be made up without limit from existing words or from parts of words; the sources of "railroad," "railway," and "aircraft" are obvious, and so are the sources of "disestablishment," first cited in 1806 and thereafter used with particular reference to the status of the Church of England. The controversy over the relations between church and state in the 19th and early 20th centuries gave rise to a chain of new words as the debate proceeded: "disestablishmentarian," "antidisestablishmentarian," "antidisestablishmentarianism." Usually, the bits and pieces of words used in this way are those found in other such combinations, but this is not always so. The technical term permafrost (terrain that never thaws, as in the Arctic) contains a bit of "permanent" probably not hitherto found in any other word.

A particular source of technical neologisms in European languages has been the words and word elements of Latin and Greek. This is part of the cultural history of western Europe, in so many ways the continuation of Greco-Roman civilization. "Microbiology" and "dolichocephalic" are words well formed according to the rules of Greek as they would be taken over into English, but no records survive of *mikrobiologia* and *dolichokephalikos* ever having been used in Ancient Greek. The same is true of Latinate creations such as "reinvestment" and "longiverbosity." The long tradition of looking to Latin and, since the Renaissance, to Greek also as the languages of European civilization, keeps alive the continuing formation of learned and scientific vocabulary in English and other European languages from these sources. The dependence on the classical languages in Europe is matched by a similar use of Sanskrit words for certain parts of learned vocabulary in some modern Indian languages (Sanskrit being the classical language of India). Such phenomena are examples of loanwords, one of the readiest sources for vocabulary extension.

Transmission of loanwords

Loanwords are words taken into a language from another language (the term borrowing is used for the process). Most obviously, this occurs when new things come into speakers' experiences as the result of contacts with speakers of other languages. This is part of the history of every language, except for one spoken by an impossibly isolated community. "Tea" from Chinese, "coffee" from Arabic, and "tomato," "potato," and "tobacco" from American Indian languages are familiar examples of loanwords designating new products that have been added to the vocabulary of English. In more abstract areas, several modern languages of India and Pakistan contain many words that relate to government, industry, and current technology taken in from English. This is the result of British rule in these countries up to indpendence and the worldwide use of English as a language of international science since then.

In general, loanwords are rapidly and completely assimilated to the prevailing grammatical and phonological patterns of the borrowing language. The German word *Kindergarten*, literally "children's garden," was borrowed into English in the middle of the 19th century to designate an informal school for young children. It is now regularly pronounced as an English word, and the plural is kindergartens (not *Kindergärten*, as in German). Occasionally, however, some loanwords retain marks of their foreign origin: examples include Latin plurals such as cacti and narcissi (as contrasted with native patterns such as cactuses and narcissuses).

Languages differ in their acceptance of loanwords. An alternative way of extending vocabulary to cope with new products is to create a descriptive compound from within one's own language. English "aircraft" and "aeroplane" are, respectively, examples of a native compound and a Greek loan creation for the same thing. English "potato" is a loan; French *pomme de terre* (literally, "apple of the earth") is a descriptive compound. Chinese is particularly resistant to loans; "aircraft," "railway," and "telephone" are translated by newly formed compounds meaning literally "fly machine," "fire vehicle," and "lightning (electricity) language."

LANGUAGE AND CONCEPTUALIZATION

The ability to speak and the ability to conceptualize are very closely linked, and the child learns both these skills together at the same time. This is not to say that thinking is no more than subvocal speech, as some behaviourists have proposed; most people can think pictorially and in simple diagrams, some to a greater degree than others, and one has the experience of responding rationally to external stimuli without intervening verbalization. But, as 18th-century thinkers saw, man's rationality developed and still goes hand in hand with his use of language, and a good deal of the flexibility of languages has been exploited in man's progressive understanding and conceptualizing of the world he lives in and of his relations with other men. Different cultures and different periods have seen this process differently developed. The anthropological linguist Edward Sapir put it well: "The 'real world' is to a large extent unconsciously built up on the language habits of the group."

Much of this lies in the irrecoverable prehistory of languages. The idea that there are still some primitive, almost "fossil" languages, embodying a very low level of conceptualization, is a vain one. All that can be said is that languages are different and that, in part, the world is seen differently through the eyes of speakers of different languages. But, in some cases, part of the lexical adaptation of a language to developing thought patterns can be followed through. Ancient Greece saw a wholly unique growth and flowering of civilization in the 1st millennium BC, which has put virtually the entire civilized world in its debt ever since. In Greek, along with the emergence of certain abstract concepts and ways of thinking, one can follow some of the changes of word meanings and the coining of new words that accompanied this. As an example, the word *dikē* originally meant "way" or "manner"; thereafter, it acquired the meaning of the right way of doing something, the right way of behaving, and finally abstract right. Its derivative *dikaiosynē*, traditionally translated "justice," became the subject of philosophical debate and analysis by the Greek philosophers and covered almost the whole range of moral obligation involved in the relations of one person with others in society. Similar debate and refinement of key terms in the various branches of thought covered by Greek philosophy can be followed through; indeed, the term philosophy is directly taken from Greek *philosophia*, a compound formed not later than the 5th century BC from *philo-* (compare *philein* "to love") and *sophia* "wisdom" to refer to abstract speculation and debate of a fundamental nature about the world and man's place in it.

Language and thought patterns

More recently, the development of the lexical resources of the languages of civilization can be observed, in one way or another, as they keep up with the scientific progress that dominates contemporary life.

An examination of the lexical structure of languages throws some light on the relations among various aspects of man's conceptualization. Spatial relations and their expression seem to lie very deep in the content of vocabulary. Words referring to time are drawn metaphorically from spatial words with great frequency: "a long/short time," "the near future," "far ahead/separated in time." Although time is a continuum, people readily divide it up into bits and record it rather as they do materials extended in space: "five years," "three months," "six seconds." This last use of vocabulary may be a particular trait of European languages and some others. An American Indian language is reported not to do this nearly so readily; it uses cardinal numbers only for discrete, countable objects. A separate class of words aligns the vocabulary of sequential time with that of intensity, so that repetition of the same activity again and again (to a European) is rather the intensification of a single activity. Certain differences in cultural attitudes and world outlook are said to accompany this kind of linguistic difference.

Expression of abstract relationships

Spatial terms are also freely used in the expression of other, more abstract relationships: "higher temperature," "higher quality," "lower expectations," "summit of a career," "far removed from any sensible course of action," "a distant relationship," "close friends," "over and above what had been said." It has been theorized that the linguistic forms most closely associated semantically with the expression of relations—case inflections in languages exhibiting this category—are originally and basically spatial in meaning. This "localist" theory, as it has been called, has been debated since the beginning of the 19th century and probably cannot be accepted as it stands, but the fact that it can be proposed and argued shows the dominant position that spatial relations hold in the conceptualization and verbalization of relations in other realms of thought.

It has been maintained that the human brain has a preference for binary oppositions, or polarities. If this is so, it will help explain the numerous pairs of related antonyms that are found: "good, bad"; "hot, cold"; "high, low"; "right, wrong"; "dark, light"; and so on. For finer discriminations, these terms can be put into more narrowly specified fields containing more than two terms taken together, but their most general use is in binary contrasts. Here, however, one term seems to represent the fundamental semantic category in question. In asking about size, one asks "How big is it?"; about weight, "How heavy is it?"; and about evaluation, "How good is it?" It is possible to ask how small, how light, or how bad something is, but such questions presuppose that the thing in mind has already been graded on the small side, on the light side, or on the bad side.

STYLE

The capacity for conceptualization possessed and developed by languages is by no means the only purpose language serves. A person's speech, supplemented by facial expression and gesture when speaker and hearer are mutually in sight, indicates and is intended to indicate a great deal more than factual information, inquiries, and requests. The fact that some of these other functions are performed by parts of a language usually mastered later by foreign learners gives rise to misinterpretation and often makes foreign speakers appear rude or insensitive when they are, in actuality, simply deploying fewer resources in the language.

Expression of emotional attitudes

Within the range of the structural and lexical possibilities of a language, speakers are able to convey their emotional attitudes and feelings toward the person or persons they are addressing and toward the subject matter of what they are saying. They are also able to conceal such feelings as one form of linguistic deception, though this is usually a harder task. These same resources are also exploited to arouse appropriate feelings and responses in others, again independently of any factual content. This is the chosen field of the propagandist, the preacher, the orator, the barrister, and the advertiser. All languages make use of intonation and voice qualities in these different ways; a person can produce and recognize the intonation and type of voice employed in coaxing, in pleading, in browbeating, and in threatening, in pleasure, and in anger, as well as those appropriate for matter-of-fact statements and the exposition of details about which the speaker has little or no emotional involvement. To describe exactly which phonetic features are brought into play is quite another matter, involving advanced competence in phonetic discrimination and analysis. This is one of the areas of speech about which all too little is currently known. Grammar and vocabulary are equally involved, though differently in each language. English speakers know the difference between "Come and give me a hand!" and "Could you possibly come and help me?"; "He's got the gift of gab" and "He is undoubtedly a fluent and persuasive speaker" are each appropriate for different occasions. By greetings and leave-takings a great deal of intended interpretation of the social relations between individuals can be expressed. Much of this is the "good manners" taught to children and expected of adults; these aspects of language behaviour vary from culture to culture, but in none are they wholly absent. It is, of course, equally possible to be deliberately bad mannered or deliberately to flout a linguistic convention or expectation, but this can be done only by knowing what is expected in the situation. The refinements of rudeness, like the refinements of politeness, insofar as the use of language is involved, require a very good knowledge of a language if it is other than one's mother tongue.

Written language is no less adapted to conveying more than just factual information, asking factual questions, and giving instructions. Intonation and tone of voice are clearly not reproducible in existing orthographic systems, but part of the skill of a novelist or a reporter is to convey these features of speech in his descriptions. Additionally, grammatical and lexical choices are available to the writer, as reading the examples above will show, and everyone knows the special artistry and techniques involved in composing written memorandums or letters if they are to achieve precisely the purpose for which they are intended.

Styles: variations within a language

These variations, written and spoken, within a language or within any dialect of a language, may be referred to as styles. Each time a person speaks or writes he does so in one or another style, deliberately chosen with the sort of considerations in mind that have just been mentioned, even though in speech the choice may often be routine. Sometimes style, especially in literature, is contrasted with "plain, everyday language." In using such plain, unmarked types of speaking or writing, however, one is no less choosing a particular style, even though it is the most commonly used one and the most neutral in that it conveys and arouses the least emotional involvement or personal feelings.

Stylistic differences are available to all mature native speakers and in literate communities to all writers, as well as to foreigners who know a second language really well. But there is undoubtedly a considerable range of skills in exploiting all the resources of a language, and, whereas all normal adults are expected to speak correctly and, if literate, to write correctly, communities have always recognized and usually respected certain individuals as preeminently skilled in particular styles, as orators, storytellers, preachers, poets, scribes, belletrists, and so forth. This is the material of literature. Once it is realized that oral literature is just as much literature as the more familiar written literature, it can be understood that there is no language devoid of its own literature.

In all languages certain forms of utterance have been considered worthy of preservation, study, and cultivation. In writing, the nature of written surfaces makes this fairly easy, though not all written material is deliberately preserved; much of it is deliberately destroyed, and, although the chance survival of inscriptions on stone or clay is of the greatest value to the archaeologist and historian, a good deal of such material was never intended to survive. Literature, on the other hand, is essentially regarded as of permanent worth. Printing and, in earlier days, the copying of manuscripts are the means of preserving written literature. In illiterate communities certain persons memorize narratives, poems, songs, prayers, ritual texts, and the like, and these are passed on, with new creations in such styles, to succeeding generations. Such skills, preservative as well as creative, are likely to be lost along with much of the surrounding culture under the impact of literacy. Here, modern technology in the guise of the tape recorder has come to the rescue, and many workers in the field of unwritten languages are recording specimens of oral literatures with transcriptions and translations while speakers having the requisite knowledge and skills are still available. A great amount of such material, however, must have been irretrievably lost from illiterate cultures before the 20th century.

Culture and literature

All languages have a literature, but different types of literature flourish in different languages and in different cultures. A warrior caste or a general respect for martial prowess fosters heroic verse or prose tales; strongly developed magical and mystery cults favour ritualistic types of oral or written literature; urban yearnings for the supposed joys of country life encourage the development of pastoral poetry, itself an outgrowth of the songs of shepherds and rural workers; and the same sense of the jadedness of city life is the best ground for the cultivation of satirical verse and prose, a form of literature probably confined largely to urban civilizations. Every language has the resources to meet these and other cultural requirements in its literature as the occasions arise, but some literary forms are more deeply involved in the structure of the language itself: this is made clear by the relative difficulty of translating certain types of literature and literary styles from one language to another. Poetry, in particular, is closely bound to the structure of the language in which it is composed, and poetry is notoriously difficult to translate from one language into another.

The special vocabularies and linguistic forms used in several games have already been mentioned. Here one may point to the widespread existence of verbal games themselves, based on the accidental features of a particular language. English-speaking children are accustomed to riddles, puns, and spelling games: "I spy with my little eye something beginning with *p*" (notice the regular formula with which this opens). These and similar word games have been found all over the world. Homer records the punning use by Odysseus of No-man (Greek *Outis*) as his name when he was about to attack Cyclops, who then roared out "No-man is killing me!" and so failed to attract any help (*Odyssey* 9:366–408). In some languages that make use of lexically distinctive tones, tone puns (words alike but for having different tones) are a form of word play.

As an intellectual challenge, the crossword puzzle in all its varieties, originally an American development early in the 20th century, has maintained and indeed greatly increased its popularity over much of the literate world that employs the Latin (Roman) alphabet. Crossword-puzzle solvers rely heavily on the relative probabilities of letter sequences in written words to suggest an answer to a partly filled line; and, depending on the particular style of the originator, crossword clues make use of many sorts of formal features in the language, among them spelling puns, spoken puns, and accidental letter sequences in words and phrases. To be able to solve a crossword puzzle in a second language shows a high degree of skill and knowledge therein.

Language and culture

It has been seen that language is much more than the external expression and communication of internal thoughts formulated independently of their verbalization. In demonstrating the inadequacy and inappropriateness of such a view of language, attention has already been drawn to the ways in which one's mother tongue is intimately and in all sorts of details related to the rest of one's life in a community and to smaller groups within that community. This is true of all peoples and all languages: it is a universal fact about language.

Language as a part of culture

Anthropologists speak of the relations between language and culture. It is, indeed, more in accordance with reality to consider language as a part of culture. "Culture" is here being used, as it is throughout this article, in the anthropological sense, to refer to all aspects of human life insofar as they are determined or conditioned by membership in a society. The fact that a man eats or drinks is not in itself cultural; it is a biological necessity that he does so for the preservation of life. That he eats particular foods and refrains from eating other substances, though they may be perfectly edible and nourishing, and that he eats and drinks at particular times of day and in certain places are matters of culture, something "acquired by man as a member of society," according to the now-classic definition of culture by the English anthropologist Sir Edward Burnett Tylor. As thus defined and envisaged, culture covers a very wide area of human life and behaviour; and language is manifestly a part, probably the most important part, of it.

Although the faculty of language acquisition and language use is innate and inherited, and there is legitimate debate over the extent of this innateness, every individual's language is "acquired by man as a member of society," along with and at the same time as other aspects of that society's culture in which he is brought up. Society and language are mutually indispensable. Language can have developed only in a social setting, however this may have been structured, and human society in any form even remotely resembling what is known today or is recorded in history could be maintained only among people speaking and understanding a language in common use.

TRANSMISSION OF LANGUAGE AND CULTURE

Language is transmitted culturally; that is, it is learned. To a lesser extent it is taught, when parents deliberately encourage their children to talk and to respond to talk, correct their mistakes, and enlarge their vocabulary. But it must be emphasized that children very largely acquire their mother tongue (*i.e.*, their first language) by "grammar construction" from exposure to a random collection of utterances that they encounter. What is classed as language teaching in school either relates to second-language acquisition or, insofar as it concerns the pupils' first language, is in the main directed at reading and writing, the study of literature, formal grammar, and alleged standards of correctness, which may not be those of all the pupils' regional or social dialects. All of what goes under the title of language teaching at school presupposes and relies on the prior knowledge of a first language in its basic vocabulary and essential structure, acquired before school age.

Transmission of culture through language

If language is transmitted as part of culture, it is no less true that culture as a whole is transmitted very largely through language, insofar as it is explicitly taught. The fact that mankind has a history in the sense that animals do not is entirely the result of language. So far as researchers can tell, animals learn through spontaneous imitation or through imitation taught by other animals. This does not exclude the performance of quite complex and substantial pieces of cooperative physical work, such as a beaver's dam or an ants' nest, nor does it preclude the intricate social organization of some species, such as bees. But it does mean that changes in organization and work will be the gradual result of mutation cumulatively reinforced by survival value; those groups whose behaviour altered in any way that increased their security from predators or from famine would survive in greater numbers than others. This would be an extremely slow process, comparable to the evolution of the different species themselves.

There is no reason to believe that animal behaviour has materially altered during the period available for the study of human history, say the last 5,000 years or so, except, of course, when man's intervention by domestication or other forms of interference has itself brought about such alterations. Nor do members of the same species differ markedly in behaviour over widely scattered areas, again apart from differences resulting from human interference. Bird songs are reported to differ somewhat from place to place within species, but there is little other evidence for areal divergence. By contrast with this unity of animal behaviour, human cultures are as divergent as are human languages over the world, and they can and do change all the time, sometimes with great rapidity, as among the industrialized nations of the 20th century.

The processes of linguistic change and its consequences will be treated below. Here, cultural change in general and its relation to language will be considered. By far the greatest part of learned behaviour, which is what culture involves, is transmitted by vocal instruction, not by imitation. Some imitation is clearly involved, especially in infancy, in the learning process, but proportionately this is hardly significant.

Through the use of language, any skills, techniques, products, modes of social control, and so on can be explained, and the end results of anyone's inventiveness can be made available to anyone else with the intellectual ability to grasp what is being said. Spoken language alone would thus vastly extend the amount of usable information in any human community and speed up the acquisition of new skills and the adaptation of techniques to changed circumstances or new environments. With the invention and diffusion of writing, this process widened immediately, and the relative permanence of writing made the diffusion of information still easier. Printing and the increase in literacy only further intensified this process. Modern techniques for almost instantaneous transmission of the written and spoken word all over the globe, together with the rapid translation services now available between the major languages of the world, have made it possible for usable knowledge of all sorts to be made accessible to people almost anywhere in the world in a very short time. This accounts for the great rapidity of scientific, technological, political, and social change in the contemporary world. All of this, whether ultimately for the good or ill of mankind, must be attributed to the dominant role of language in the transmission of culture.

Diffusion of knowledge through writing and printing

LANGUAGE AND SOCIAL DIFFERENTIATION AND ASSIMILATION

The part played by variations within a language in differentiating social and occupational groups in a society has already been referred to above. In language transmission this tends to be self-perpetuating unless deliberately interfered with. Children are in general brought up within the social group to which their parents and immediate family circle belong, and they learn the dialect and speaking styles of that group along with the rest of the subculture and behavioral traits and attitudes that are characteristic of it. This is a largely unconscious and involuntary process of acculturation, but the importance of the linguistic manifestations of social status and of social hierarchies is not lost on aspirants for personal advancement in stratified societies. The deliberate cultivation of an appropriate dialect, in its lexical, grammatical, and phonetic features, has been the self-imposed task of many persons wishing "to better themselves" and the butt of unkind ridicule on the part of persons already feeling themselves secure in their social status or unwilling to attempt any change in it. Much of the comedy in George Bernard Shaw's *Pygmalion* turns on Eliza's need to unlearn her native Cockney if she is to rise in the social scale. Conversely, it is readily apparent today that middle class people, mostly adolescents, who for some reason want to "opt out" of the social group of their parents make every effort to abandon the distinctive aspects of the social dialect that would mark them, along with dress and general behaviour, as members of a group whose mores they are, at least temporarily, affecting to reject. Culturally and subculturally determined taboos play a part in all this, and persons desirous of moving up or down in the social scale have to learn what words to use and what words to avoid if they are to be accepted and to "belong" in their new position. All through the ages, a good part of the material for "comedies of manners" has come from the social role of language variation within a society.

The same considerations apply to changing one's language as to changing one's dialect. Language changing is harder for the individual and is generally a rarer occurrence, but it is likely to be widespread in any mass immigration movement. In the 19th and early 20th centuries, the eagerness with which immigrants and the children of immigrants from continental Europe living in the United States learned and insisted on speaking English is an illustration of their realization that English was the linguistic badge of full membership in their new homeland at the time when the country was proud to consider itself as the melting pot in which people of diverse linguistic and cultural origins would become citizens of a unified community.

Special vocabularies of games, trades, professions

The same sort of self-perpetuation, in the absence of deliberate rejection, operates in the special languages of games and of trades and professions (these are in the main concerned with special vocabularies). Game learners, apprentices, and professional students learn the locutions together with the rest of the game or the job. The specific words and phrases occur in the teaching process and are observed in use, and the novice is only too eager to display an easy competence with such phraseology as a mark of his full membership of the group; *e.g.*, golfers are keen to talk of birdies, fairways, and slicing.

Languages and variations within languages play both a unifying and a diversifying role in human society as a whole. Language is a part of culture, but culture is a complex totality containing many different features, and the boundaries between cultural features are not clear-cut, nor do they all coincide. Physical barriers such as oceans, high mountains, and wide rivers constitute impediments to human intercourse and to culture contacts, though modern technology in the fields of travel and communications make such geographical factors of less and less account. More potent for much of the 20th century were political restrictions on the movement of people and of ideas, such as divided western Europe from formerly Communist eastern Europe; the frontiers between these two political blocs represented much more of a cultural dividing line than any other European frontiers.

The distribution of the various components of cultures differs, and the distribution of languages may differ from that of nonlinguistic cultural features. This results from the varying ease and rapidity with which changes may be acquired or enforced and from the historical circumstances responsible for these changes. In mid- to late-20th-century Europe, as the result of World War II, a major political and cultural division had cut across an area of relative linguistic unity in East and West Germany. It is significant, however, that differences of vocabulary and usage were soon noticeable in the German speech from each side, overlying earlier differences attributed to regional dialects; although the two countries were unified in 1990, the east-west division may have marked a definite dialect boundary within the German language as well.

THE CONTROL OF LANGUAGE FOR CULTURAL ENDS

Second-language learning. Language, no less than other aspects of human behaviour, is subject to purposive interference. When people with different languages need to communicate, various expedients are open to them, the most obvious being second-language learning and teaching. This takes time, effort, and organization, and, when more than two languages are involved, the time and effort are that much greater. Most people are monolingual, and those with a working knowledge of three or four languages are much fewer than those with a competence in just one second language. Other expedients may also be applied. Ad hoc pidgins for the restricted purposes of trade and administration were mentioned above. Tacit or deliberate agreements have been reached whereby one language is chosen for international purposes when speakers of several different languages are involved. In the Roman Empire, broadly, the western half used Latin as a lingua franca, and the eastern half used Greek. In western Europe during the Middle Ages, Latin continued as the international language of educated people, and Latin was the second language taught in schools. Later, the cultural, diplomatic, and military reputation of France made French the language of European diplomacy. This use of French as the language of international relations persisted until the 20th century. At important conferences among representatives of different nations, it is usually agreed which languages shall be officially recognized for registering the decisions reached; and the provisions of treaties are interpreted in the light of texts in a limited number of languages, those of the major participants.

International and diplomatic languages

Since World War II the dominance of the English-speaking peoples in science and technology and in international commerce has led to the recognition of English as the major international language in the world of practical affairs, with more and more countries making English the first foreign language to be taught and thus producing a vast expansion of English-language-teaching programs all over the world. Those whose native language is English do not sufficiently realize the amount of effort, by teacher and learner alike, that is put into the acquisition of a working knowledge of English by educated first speakers of other languages.

As an alternative to the recognition of particular natural languages as international in status, attempts have been made to invent and propagate new and genuinely international languages, devised for the purpose. Of these, Esperanto, invented by the Polish-Russian doctor L.L. Zamenhof in the 19th century, is the best known. Such languages are generally built up from parts of the vocabulary and grammatical apparatus of the better known existing languages of the world. The relationship between the written letter and its pronunciation is more systematic than with many existing orthographies (English spelling is notoriously unreliable as an indication of pronunciation), and care is taken to avoid the grammatical irregularities to which all natural languages are subject and also to avoid sounds found difficult by many speakers (*e.g.*, the English *th* sounds, which most Europeans, apart from English speakers, dislike). These artificial languages have not made much progress, though an international society of Esperanto speakers does exist.

Nationalistic influences on language. Deliberate interference with the natural course of linguistic changes and the distribution of languages is not confined to the facilitating of international intercourse and cooperation. Language as a cohesive force for nation-states and for linguistic groups within nation-states has for long been manipulated for political ends. Multilingual states can exist and prosper; Switzerland is a good example. But linguistic rivalry and strife can be disruptive. Language riots have occurred in Belgium between French and Flemish speakers and in parts of India between rival vernacular communities. A language can become or be made a focus of loyalty for a minority community that thinks itself suppressed, persecuted, or subjected to discrimination. The French language in Canada in the mid-20th century is an example. In the 19th and early 20th centuries Irish Gaelic came to symbolize Irish patriotism and Irish independence from Great Britain. Since independence, government policy continues to insist on the equal status of English and Irish in public notices and official documents, but, despite such encouragement and the official teaching of Irish in the state schools, a main motivation for its use and study has disappeared, and the language is giving ground to English under the international pressures referred to above.

Language as a cohesive political force

For the same reasons, a language may be a target for attack or suppression, if the authorities associate it with what they consider a disaffected or rebellious group or even just a culturally inferior one. There have been periods when American Indian children were forbidden to speak a language other than English at school and when pupils were not allowed to speak Welsh in British state schools in Wales. Both these prohibitions have been abandoned. Since the Spanish Civil War of the 1930s Basque speakers have been discouraged from using their language in public, as a consequence of the strong support given by the Basques to the republican forces. Interestingly, on the other side of the Franco-Spanish frontier, French Basques are positively encouraged to keep their language in use, if only as an object of touristic interest and consequent economic benefit to the area.

Translation. So far, some of the relatively large-scale effects of culture contacts on languages and on dialects within languages have been surveyed. A continuous concomitant of contact between two mutually incomprehensible tongues and one that does not lead either to suppression or extension of either is translation. As soon as two speakers of different languages need to converse, translation is necessary, either through a third party or directly.

Before the invention and diffusion of writing, translation was instantaneous and oral; persons professionally specializing in such work were called interpreters. In predominantly or wholly literate communities, translation is thought of as the conversion of a written text in one language into a written text in another, though the modern emergence of the simultaneous translator or professional interpreter at international conferences keeps the oral side of translation very much alive.

Oral and written translation

The tasks of the translator are the same whether the material is oral or written, but, of course, translation between written texts allows more time for stylistic adjustment and technical expertise. The main problems have been recognized since antiquity and were expressed by St. Jerome, translator of the famed Latin Bible, the Vulgate, from the Hebrew and Greek originals. Semantically, these problems relate to the adjustment of the literal and the literary and the conflicts that so often occur between an exact translation of each word, as far as this is possible, and the production of a whole sentence or even a whole text that conveys as much of the meaning of the original as can be managed. These problems and conflicts arise because of factors already noticed in the use and functioning of language: languages do not operate in isolation but within and as part of cultures, and cultures differ from each other in various ways. Even between the languages of communities whose cultures are fairly closely allied, there is by no means a one-to-one relation of exact lexical equivalence between the items of their vocabularies.

In their lexical meanings, words acquire various overtones and associations that are not shared by the nearest corresponding words in other languages; this may vitiate a literal translation. The English author and theologian Ronald Knox has pointed to the historical connections of the Greek *skandalon* "stumbling block, trap, or snare," inadequately rendered by "offense," its usual New Testament translation. In modern times translators of the Bible into the languages of peoples culturally remote from Europe are well aware of the difficulties of finding a lexical equivalent for "lamb," when the intended readers, even if they have seen sheep and lambs, have no tradition of blood sacrifice for expiation nor long-hallowed associations of lambs with lovableness, innocence, and apparent helplessness. The English word uncle has, for various reasons, a cozy and slightly comic set of associations. The Latin poet Virgil uses the words *avunculus Hector* in a solemn heroic passage of the Aeneid (Book III, line 343); to translate this by "uncle Hector" gives an entirely unsuitable flavour to the text.

The translation of poetry, especially into poetry, presents very special difficulties, and the better the original poem, the harder the translator's task. This is because poetry is, in the first instance, carefully contrived to express exactly what the poet wants to say. Second, to achieve this end, the poet calls forth all the resources of the language in which he is writing, matching the choice of words, the order of words, and grammatical constructions, as well as phonological features peculiar to the language in metre, perhaps supplemented by rhyme, assonance, and alliteration. The available resources differ from language to language; English and German rely on stress-marked metres, but Latin and Greek used quantitative metres, contrasting long and short syllables, while French places approximately equal stress and length on each syllable. The translator must try to match the stylistic exploitation of the particular resources in the original language with comparable resources from his own. Because lexical, grammatical, and metrical considerations are all interrelated and interwoven in poetry, a satisfactory literary translation is usually very far from a literal word for word rendering. The more the poet relies on language form, the more embedded his verses are in that particular language, and the harder they are to translate adequately. This is especially true with lyrical poetry in several languages, with its wordplay, complex rhymes, and frequent assonances.

Poetry translation

At the other end of the translator's spectrum, technical prose dealing with internationally agreed scientific subjects is probably the easiest type of material to translate, because cultural unification (in this respect), lexical correspondences, and stylistic similarity already exist in this type of usage in the languages most commonly involved, to a higher degree than in other fields of discourse.

Significantly, it is this last aspect of translation to which mechanical and computerized techniques are being applied with some prospects of limited success. Machine translation, whereby, ultimately, a text in one language could be fed into a machine to produce an accurate translation in another language without further human intervention, has been largely concentrated on the language of science and technology, with its restricted vocabulary and overall likeness of style, for both linguistic and economic reasons. Attempts at machine translation of literature have been made, but success in this field, more especially in the translation of poetry, seems very remote at present.

Translation as an art

Translation on the whole is an art, not a science. Guidance can be given and general principles can be taught, but after that it must be left to the individual's own feeling for the two languages concerned. Almost inevitably, in a translation of a work of literature something of the author's original intent must be lost; in those cases in which the translation is said to be a better work than the original, an opinion sometimes expressed about the English writer Edward Fitzgerald's "translation" of *The Rubáiyát of Omar Khayyám*, one is dealing with a new, though derived, work, not just a translation. The Italian epigram remains justified: *Traduttore traditore* "The translator is a traitor."

Messages and codes. Translation serves to extend the communicative value of a text. Sometimes people want to restrict it. Confidential messages, spoken and written, require for their efficacy that they be known to and understood by only the single person or the few persons to whom they are addressed. Such are diplomatic exchanges, operational messages in wartime, and some transmissions of commercial information. Protection of written messages from interception has been practiced for many centuries. Recent developments in telegraphy and telephony have made protection against unauthorized reception more urgent, whether of texts transmitted as speech or as series of letters of the alphabet. Scrambling of telephony is a common expedient; the wave frequencies through which the sounds are to be transmitted are altered at the source so as to be unrecognizable and then reconverted by the intended recipient's receiver. Codes and ciphers (cryptography) are of much longer standing in the concealment of written messages, though their techniques are being constantly developed. Such gains are, of course, countered by developments in the techniques of decipherment and decoding (as distinct from getting hold of the key to the system in use). An important by-product of such techniques has been the reading and interpretation of inscriptions written in otherwise unknown languages or unknown writing systems for which no translation exists. The recent, very significant decipherment of the Linear B script and its recognition as Mycenaean Greek, an early Greek dialect written in a form of orthography quite distinct from the later classical Greek alphabet, was first achieved by the application of cryptographic "code cracking" methods (see also CRYPTOLOGY).

LANGUAGE LEARNING

Every physiologically and mentally normal person has learned the main structure and basic vocabulary of his mother tongue by the end of childhood. It has been pointed out that the process of first-language acquisition as a spoken medium of communication is largely achieved from random exposure. There is legitimate controversy, however, over the nature and extent of the positive contribution that the human brain brings to the activity of grammar construction, the activity by which the child develops an indefinitely creative competence from the finite data that make up his actual experience of the language. Creativity is what must be stressed as the product of first-language acquisition. By far the greater number of all the sentences anyone hears and utters during his lifetime are new; that is, they have not occurred before in his personal experience. But individuals find no difficulty at all in understanding at once almost everything they hear nor for the most part in producing sentences to suit the requirements of every situation. This very ease of creativity in man's linguistic competence makes it hard to realize its extent. The only regularly reproduced sentences in most speakers' experience are the stereotyped forms of greeting and leave-taking and certain formalized responses to recurrent situations, such as shopping, cooperative activities in repetitive jobs, the stylized parts of church services, and the like.

Role of the brain in language learning

Yet, despite this really immense achievement that the progressive mastery of one's first language constitutes, it arouses no comment and attracts no credit. It is simply part of what is expected of one in growing up. Different people may be singled out for praise in certain uses of their language, as good public speakers, authors, poets, tellers of tales, and solvers of puzzles, but not just as speakers. The credit that some individuals acquire in certain communities for "speaking correctly" is a different matter, usually the result of speaking as one's mother tongue a prestigious standard dialect among people most of whom speak another, less favoured one.

Bilingualism. The learning of a second and of any subsequently acquired language is quite a different matter. Except for one form of bilingualism (see below), it is a deliberate activity undertaken when one has already nearly or fully acquired the basic structure and vocabulary of one's first language. Of course, many people never do master significantly more than their own first language. It is only in encountering a second language that one realizes how complex language is and how much effort must be devoted to subsequent acquisition. It has been said that the principal obstacle to learning a language is knowing one already, and it may also be that the faculty of grammar construction exhibited in childhood is one that is gradually lost as childhood recedes.

Ability to learn additional languages

Whereas every normal person masters his mother tongue with unconscious ease, people vary in their ability to learn additional languages, just as they vary in other intellectual activities. Situational motivation, however, appears to be by far the strongest influence on the speed and apparent ease of this learning. The greatest difficulty is experienced by those who learn because they are told to or are expected to, without supporting reasons that they can justify. Given a motive other than external compulsion or expectation, the task is achieved much more easily (this, of course, is an observation in no way confined to language learning). In Welsh schools it is found that English children make slower progress in Welsh when their only apparent reason for learning Welsh is that there are Welsh classes. Welsh children, on the other hand, make rapid progress in English, the language of most further education, the newspapers, most television and radio, most of the better paid jobs, and of any job outside Welsh-speaking areas. Similar differences in motivation have accounted for the excellent standard of English, French, and German acquired by educated persons in the Scandinavian countries and in Holland, small countries whose languages, being spoken by relatively few foreigners, are of little use in international communication. This attainment may be compared with the much poorer showing in second-language acquisition among comparably educated persons in England and America, who have for long been able to rely on foreigners accommodating to their ignorance by speaking and understanding English.

It is often held that children brought up bilingually in places in which two languages are regularly in use are slower in schoolwork than comparable monolingual children, as a greater amount of mental effort has to be expended in the mastery of two languages. This is by no means proved; and, because much of a child's language acquisition takes place in infancy and in the preschool years, it does not represent an effort in the way that consciously learning a language in school does, and indeed it probably occupies a separate part of the child's mental equipment. The question of speed of general learning by bilinguals and monolinguals must be left open. It is quite a separate matter from the job of learning, by teaching at home or in school, to read and write in two languages; this undoubtedly is more of a labour than the acquisition of monolingual literacy.

Types of bilingualism

Two types of bilingualism have been distinguished, according to whether the two languages were acquired from the simultaneous experience of the use of both in the same circumstances and settings or from exposure to each language used in different settings (an example of the latter is the experience of English children living in India during the period of British ascendancy there, learning English from their parents and an Indian language from their nurses and family servants). However acquired, bilingualism leads to mutual interference between the two languages; extensive bilingualism within a community is sometimes held partly responsible for linguistic change (see below). Interference may take place in pronunciation, in grammar, and in the meanings of words. Bilinguals often speak their two languages each with "an accent"; *i.e.*, they carry into each certain pronunciation features from the other. The German word order in "He comes tomorrow home" has been reported as an example of grammatical interference; and in Candian French the verb *introduire* has acquired from English the additional meaning "introduce, make acquainted" (which in metropolitan French is *présenter*).

Literacy. The acquisition of literacy is something very different from the acquisition of one's spoken mother tongue, even when the same language is involved, as it usually is. Both skills, speaking and writing, are learned skills, but there the resemblance ends. The child learns his first language at the start involuntarily and mostly unconsciously from random exposure, even if no attempts at teaching are made. Literacy is deliberately taught and consciously and deliberately learned. There is current debate on the best methods and techniques for teaching literacy in various social and linguistic settings. Literacy is learned through speech, by a person already possessed of the basic structure and vocabulary of his language.

Such facts should be very obvious, but the now-accepted, though fairly recent, standard of near-universal literacy in technologically advanced countries, along with the fact that in second-language learning one usually acquires speech and writing skills at the same time, tends to bring these two parts of language learning under one head. Literacy is manifestly a desirable attainment for all communities, though not necessarily in all languages. It must be borne in mind that there are many distinct languages spoken in the world today by fewer than 1,000 or 500 or even 50 persons. The capital investment in literacy, including teaching resources, teacher time and training, printing, publications, and so forth, is vast, and it can be economically and socially justified only when applied to languages spoken and likely to continue to be spoken by substantial numbers over a wide area.

Effects of literacy

Literacy is in no way necessary for the maintenance of linguistic structure or vocabulary, though it does enable people to add words from the common written stock in dictionaries to their personal vocabulary very easily. It is worth emphasizing that until relatively recently in human history all languages were spoken by illiterate speakers and that there is no essential difference as regards pronunciation, structure, and complexity of vocabulary between spoken languages that have writing systems used by all or nearly all their speakers and the languages of illiterate communities.

Literacy has many effects on the uses to which language may be put; storage, retrieval, and dissemination of information are greatly facilitated, and some uses of language, such as philosophical system building and the keeping of detailed historical records, would scarcely be possible in a totally illiterate community. In these respects the lexical content of a language is affected, for example, by the creation of sets of technical terms for philosophical writing and debate. Because the permanence of writing overcomes the limitations of auditory memory span imposed on speech, sentences of greater length can easily occur in writing, especially in types of written language that are not normally read aloud and that do not directly represent what would be spoken. An examination of some kinds of oral literature, however, reveals the ability of the human brain to receive and interpret spoken sentences of considerable grammatical complexity.

In relation to pronunciation, writing does not prevent the historical changes that occur in all languages. Part of the apparent irrationality of English spelling, such as is found also in some other orthographies, lies just in the fact that letter sequences have remained constant while the sounds represented by them have changed. For example, the *gh* of "light" once stood for a consonant sound, as it still does in the word as pronounced in some Scots dialects; and the *k* of "knave" and "knight" likewise stood for an initial *k* sound (compare the related German words *Knabe* and *Knecht*). A few relatively uncommon words, including some proper names, are reformed phonetically, specifically to bring their pronunciation more in line with their spelling. Spelling prounciations, as these are called, are a product of general literacy. In London, the pronunciation of "St. Mary Axe" as if it were spelled "Simmery Axe" is now decidedly old-fashioned. "St. John" and "St. Clair" survive as proper names with their old pronunciations, in the latter case helped by the presence of the alternative spelling "Sinclair."

WRITTEN LANGUAGE

Historically, culturally, and in the individual's life, writing is subsequent to speech and presupposes it. Aristotle expressed the relation thus: "Speech is the representation of the experiences of the mind, and writing is the representation of speech" (*On Interpretation*). But it is not as simple as this would suggest. Alphabetic writing, in which, broadly, consonant and vowel sounds are indicated by letters in sequence, is the most widespread system in use today, and it is the means by which literacy will be disseminated, but it is not the only system, nor is it the earliest.

Evolution of writing systems. Writing appears to have been evolved from an extension of picture signs: signs that directly and iconically represented some thing or action and then the word that bore that meaning. Other words or word elements not readily represented pictorially could be assigned picture signs already standing for a word of the same or nearly the same pronunciation, perhaps with some additional mark to keep the two signs apart. This sort of device is used in children's word puzzles, as when the picture of a berry is used to represent, say, the second half of the name Canterbury. This opens the way for what is called a character script, such as that of Chinese, in which each word is graphically represented by a separate individual symbol or character or by a sequence of two or more such characters. Writing systems of this sort have appeared independently in different parts of the world.

Character scripts

Chinese character writing has for many centuries been stylized, but it still bears marks of the pictorial origin of some characters. Chinese characters and the characters of similar writing systems are sometimes called ideograms, as if they directly represented thoughts or ideas. This is not so. Chinese characters stand for Chinese words or, particularly as in modern Chinese, bits of words; they are the symbolization of a particular language, not a potentially universal representation of thought. The ampersand (&) sign, standing for "and" in English printing, is a good isolated example of a character used in an alphabetic writing system.

Character writing is laborious to learn and imposes a burden on the memory. Alternatives to it, in addition to alphabetic writing, include scripts that employ separate symbols for the syllable sequences of consonants and vowels in a language, with graphic devices to indicate consonants not followed by a vowel. The Devanagari script, in which classical Sanskrit and modern Hindi are written, is of this type, and the Mycenaean writing system, a form of Greek writing in use in the 2nd millennium BC and quite independent of the later Greek alphabet, was syllabic in structure. Japanese employs a mixed system, broadly representing the roots of words by Chinese characters (the Japanese learned writing from the Chinese in and after the 5th century AD) and the inflectional endings by syllable signs. These syllable signs are an illustration of the way in which a syllabic script can develop from a character script: certain Chinese characters were selected for their sound values alone and, reduced in size and complexity, have been standardized as signs of a particular consonant and vowel sequence or of a single vowel sound.

The Greek alphabet came from the Phoenician script, a

Greek alphabet

syllabic-type writing system that indicated the consonant sounds. By a stroke of genius, a Greek community decided to employ certain consonantal signs to which no consonant sound corresponded in Greek as independent vowel signs, thus producing an alphabet, a set of letters standing for consonants and vowels. The Greek alphabet spread over the ancient Greek world, undergoing minor changes. From a Western version sprang the Latin (Roman) alphabet. Also derived from the Greek alphabet, the Cyrillic alphabet was devised in the 9th century AD by a Greek missionary, St. Cyril, for writing the Slavic languages.

Spelling. Alphabetic writing is not and cannot be an exact representation of the sequence of sounds or even of the sequence of distinctive sounds in the spoken forms of words and sentences. "Consonant" and "vowel" mean different things when applied to letters and to sounds, though there is, of course, much overlap. The *y* at the beginning of "yet" stands for a consonant sound; at the end of "jetty" it stands for a vowel sound. In "thick" and "thin" the sequence *th* represents a single sound, not a *t* sound followed by an *h* sound. In "kite" the *e* represents no sound directly but distinguishes the vowel between *k* and *t* from the vowel in "kit." These disharmonies arise from a number of causes. Economy in the use of letters is one factor. In addition, spoken forms are always changing over the centuries, whereas writing, particularly since the invention of printing, is very conservative. At one time the *e* at the end of words such as "kite" did stand for a vowel sound. This sound was lost between the 14th and 16th centuries, a time when other changes in the pronunciation of such words also occurred. The notorious *ough* spellings in English, standing for different sounds and sound sequences in "rough," "cough," "dough," "plough," "ought," and other such words, have arisen from historical changes that have driven spelling and pronunciation further apart.

This, of course, does not mean that spelling reforms are out of the question. Spelling reform has been talked of in relation to English for many centuries without much effect; but in some countries—for example, Norway and Holland—official action has prescribed certain reforms to be made, and these have then been taught in school and have gradually found their way into printed works. The sheer volume of printed matter preserved for use and consultation in the modern world adds much weight against the convenience otherwise accruing from reforms designed to correct the historically produced disharmonies between spelling and pronunciation.

Representing grammatical distinctions in spelling

Moreover, it is not always most useful for spellings to represent exactly the sound sequences in a word and nothing else; this is the task for which phoneticians have devised transcriptions. As far as the sounds themselves are concerned, the plural signs of "cats," "dogs," and "horses" are different: the final sound of "cats" is like the initial sound of "sink," that of "dogs" like the initial sound of "zinc," and the plural of "horse" is indicated by a sound sequence rather like that in "is." But they are all indicated in writing by one and the same letter and always have been, because only one grammatical distinction, that of singular as against plural, is involved, and at this point in the language the actual differences in the sounds, important elsewhere, are irrelevant.

Letters, insofar as they stand for sounds, stand for consonants and vowels. But other sound features are involved in languages. In English words the location of the stress is important, and the words "import" as a noun and "import" as a verb are distinguished by this alone. All languages make use of sequences of rises and falls in pitch, called intonation, as part of spoken communication. These phenomena are unrepresented in orthography except for certain punctuation marks such as ? and ! and sometimes by italicization and underlining.

This is not a weakness in orthography. Writing is normally intended to be read and when necessary read aloud by people who already know the language and are therefore able to supply from their own competence the required detail. For specific purposes such as foreign-language teaching, as well as for the specific study of pronunciation and speech sounds in phonetics and phonology, various forms of transcription have been devised to indicate unambiguously by written signs the precise form of the spoken utterance, without regard to other considerations.

Written versus spoken languages. For these reasons one should distinguish the grammar of a written language (*e.g.*, written English) from the grammar of the corresponding spoken language (spoken English). The two grammars will be very similar, and they will overlap in most places; but the description of spoken English will have to take into account the grammatical uses of features such as intonation, largely unrepresented in writing, and the description of written English must deal adequately with the greater average length of sentences and some different syntactic constructions and word forms characterizing certain written styles but almost unknown in ordinary speech (*e.g.*, "whom" as the objective form of "who").

Ancient, or dead, languages

In studying ancient (dead) languages one is, of course, limited to studying the grammar of their written forms and styles, as their written records alone survive. Such is the case with Latin, Ancient Greek, and Sanskrit (Latin lives as a spoken language in very restricted situations, such as the official language of some closed religious communities, but this is not the same sort of Latin as that studied in classical Latin literature; Sanskrit survives also as a spoken language in similarly restricted situations in a few places in India). Scholars may be able to reconstruct something of the pronunciation of a dead language from historical inferences and from descriptions of its pronunciation by authors writing when the language was still spoken. They know a good deal about the pronunciation of Greek and Latin and a great deal about the pronunciation of Sanskrit, because ancient Indian scholars left a collection of extremely detailed and systematic literature on its pronunciation. But this does not alter the fact that when one teaches and learns dead languages today, largely for their literary value and because of the place of the communities formerly speaking them in our own cultural history, one is teaching and learning the grammar of their written forms. Indeed, despite what is known about the actual pronunciation of Greek and Latin, Europeans on the whole pronounce what they read in terms of the pronunciation patterns of their own languages.

Under present conditions, with universal literacy either an accepted fact or an accepted target, it is assumed that, wherever it is convenient or useful, writing may be employed for any purpose for which speech might have been used and by all sections of the community. This has not always been so. Literacy was until the 19th century the privilege of the few. In other periods and cultures, writing was the preserve of certain defined groups, such as the priesthood and the official class, and it was restricted to certain purposes, such as the annals of important events, genealogical tables, and records of inventories of things and persons. It is highly probable that writing first developed for particular types of use by particular groups of specialists within communities and subsequently, because of its obvious utility, spread outside these limits.

For further accounts of writing systems in greater detail, see WRITING.

Linguistic change

Every language has a history; and, as in the rest of human culture, changes are constantly taking place in the course of the learned transmission of a language from one generation to another. This is just part of the differences between human culture and animal behaviour. Languages change in all their aspects, in their pronunciation, word forms, syntax, and word meanings (semantic change). These changes are mostly very gradual in their operation, becoming noticeable only cumulatively over the course of several generations. But, in some areas of vocabulary, particular words closely related to rapid cultural change are subject to equally rapid and therefore noticeable changes within a generation or even within a decade. In the 20th century the vocabulary of science and technology is an outstanding example. The same is also true of those parts of vocabulary that are involved in fashionable slangs and jargons, whose raison d'être in promoting group, particu-

Changes in vocabulary

larly age-group, solidarity depends on their being always fresh and distinctive. Old slangs date, as any reading of a novel or visit to a film more than 10 years old is apt to show. The rapid obsolescence of young people's slangs is equally to be seen in the unsuccessful efforts of some well-intentioned older persons who vainly attempt to cultivate the speech styles of present-day youth groups in a misdirected attempt to bridge "the generation gap" (this last phrase is an example of mid-20th-century pseudoscientific slang).

DIVERSIFICATION OF LANGUAGES

Changes through time. In the structural aspects of language, their pronunciation and grammar, and in vocabulary less closely involved in rapid cultural movement, the processes of linguistic change are best observed by comparing written records of a language over extended periods. This is most readily seen by English speakers through setting side by side present-day English texts with 18th-century English, the English of the Authorized Version of the Bible, Shakespearean English, Chaucer's English, and the varieties of Old English (Anglo-Saxon) that survive in written form. Noticeably, as one goes back in time, the effort required in understanding increases, and, while people do not hesitate to speak of "Shakespearean English," they are more doubtful about Chaucer, and for the most part Old English texts are as unintelligible to a modern English speaker as, for example, texts in German. It is clear that the differences involved include word meanings, grammar, and, so far as this can be reconstructed, pronunciation.

Similar evidence, together with what is known of the cultural history of the peoples concerned, makes clear the continuous historical connections linking French, Spanish, Portuguese, Italian, and Romanian with the spoken ("vulgar") Latin of the western Roman Empire. This group constitutes the Romance subfamily of languages and is an example of how, as the result of linguistic change over a wide area, a group of distinct, though historically related, languages comes into being.

In the transmission of a language from parent to child, slight deviations in all aspects of language use occur all the time, and as the child's speech contacts widen he confronts a growing range of slight differences in personal speech forms, some of them correlating with social or regional differences within a community, these speech differences themselves being the results of the transmission process. As a consequence, the child's speech comes to differ slightly from that of his parents' generation. In urbanized communities an additional factor is involved: children have been shown to be effectively influenced by the speech habits of their peer groups once they have made contacts with them in and out of school.

Cumulative changes in languages

Such changes, though slight at the time, are progressively cumulative. Since ready intercommunication is a primary purpose of language, as long as a community remains unitary, with strong central direction and a central cultural focus, such changes will not go beyond the limits of intercomprehensibility. But in more scattered communities and in larger language areas, especially when cultural and administrative ties are weakened and broken, these cumulative deviations in the course of generations give rise to wider regional differences. Such differences take the form of dialectal differentiation as long as there is some degree of mutual comprehension but eventually result in the emergence of distinct languages. This is what happened in the history of the colloquial Latin of the western Roman Empire, and it can be assumed that a similar course of events gave rise to the separate Germanic languages (English, German, Dutch, Danish, Norwegian, Swedish, and some others), though in this family the original unitary language is not known historically but inferred as "Common Germanic" or "Proto-Germanic" and tentatively assigned to early in the 1st millennium BC as the period before separation began.

This is how language families have developed. Most but not all of the languages of Europe belong to the Indo-European family, so-called because in addition it includes the classical Indian language Sanskrit and most of the modern languages of northern India and Pakistan. It includes as subfamilies the two families just mentioned, Romance and Germanic, and several others. It is assumed that the subfamilies, and from them the individual languages of the Indo-European family, are ultimately derived from a unitary language spoken somewhere in eastern Europe or western Asia (its exact location is still under debate), perhaps 5,000 years ago. This unitary language has itself been referred to as "Indo-European," "Proto-Indo-European," the "common parent language," or the "original language" (*Ursprache*) of the family. But it must be emphasized that, whatever it may have been like, it was just one language among many and of no special status in itself. It was certainly in no way the original language of mankind or anything like it. It had its own earlier history, of which virtually nothing can be inferred, and it was, of course, very recent in relation to the time span of human language itself. What is really special about such "parent" or "proto-" languages is that they represent the farthest point to which our available techniques and resources enable us to reconstruct the prehistory of our attested and living languages. Similarly constituted families of languages derived from inferred common sources have been established for other parts of the world; for example, Altaic, covering Turkish and several languages of Central Asia, and Bantu, containing many of the languages of central and southern Africa. For further details of these and other language families see LANGUAGES OF THE WORLD.

"Parent" or "proto-" languages

If enough material in the form of written records from past ages were available, it would be possible to group all the world's languages into historically related families. In addition, an answer could perhaps be posited to the question of whether all languages are descended from a single original language or whether languages emerged independently among several groups of early peoples (the rival theories of monogenesis and polygenesis, a controversy more confidently disputed in the 19th century than today). In actual fact, written records, when they are available, go back only a fraction of the time in which human speech has been developed and used, and over much of the globe written records are nonexistent. In addition, there are no other linguistic fossils comparable to the fossils of geological prehistory. This means that the history and prehistory of languages will not be able to go back more than a few thousand years BC and will be much more restricted in language areas in which few or no written records are available, as in much of Africa and in South America. Many languages will remain not related with certainty to any family. Nevertheless, the methods of historical linguistics, involving the precise and systematic comparison of word forms and word meanings (see further LINGUISTICS), have produced remarkable results in establishing language families on the same basis as Indo-European was established, in far less favourable fields. But any attempt by these means to get back to "the origin of language" or to reconstruct man's original language, if indeed there was one, is quite beyond the reach of science and will remain so.

Changes through geographical movement. The fundamental cause of linguistic change and hence of linguistic diversification is the minute deviations occurring in the transmission of speech from one generation to another. But other factors contribute to the historical development of languages and determine the spread of a language family over the world's surface. Population movements naturally play a large part, and movements of peoples in prehistoric times carried the Indo-European languages from a relatively restricted area into most of Europe and into northern India, Persia, and Armenia. But language and race are by no means the same thing, and the spread of the Indo-European languages resulted, in the main, from the imposition of one of them on the earlier population of the territories occupied. In the historical period, within Indo-European, the same process can be seen at work in the western Roman Empire. Latin superseded the earlier, largely Celtic languages of the Iberian Peninsula and of Gaul (France) not through population replacement (the number of Roman soldiers and settlers in the empire was never large) but through the abandonment of these languages by the inhabitants over the generations as they

Population movements

found in Latin the language of commerce, civilization, law, literature, and social prestige.

Conquest does not always lead to the supersession of a language. Greek survived centuries of Turkish rule and indeed remained a focus of national feeling, as has happened elsewhere in history. Much depends on the various circumstances and on the mutual attitudes of those involved; what must be kept quite clear is the difference between movements of peoples and the spread of languages. When linguistically homogeneous people enter and occupy a virtually empty area, as with most of Australia, the two movements coincide.

Languages do not just spread and compete with each other for territorial use. They are in constant contact, and every language bears evidence of this throughout its history. Modern Greek is full of words of Turkish origin, despite efforts made at various times since independence to purify the language by official action. The Norman Conquest and a period in which French was the language of the ruling class in England effected great changes on English and contributed a very substantial number of French words to English vocabulary; hence the quantity of near synonymous pairs available today: "begin, commence"; "end, finish"; "kingly, royal"; "fight, combat"; and so on.

Tendencies against change. These historical processes take place without any direct volition on the part of speakers as regards the language itself. Latin was learned as part of personal advancement, not for its own sake. Loans were incorporated almost without their being noticed, along with the concomitant cultural changes and innovations. Deliberate action directly related to a language does occur. The creation of pidgins involves some degree of linguistic consciousness on the part of their first users. More deliberate, however, have been various attempts at preserving the purity of a language, at least for some uses, or at arresting the processes of change. The care bestowed on the preservation of the Sanskrit used in religious ritual in ancient India and recent attempts to free Modern Greek from much of its Turkish vocabulary have already been noticed. For a period, under Nazi rule, efforts were made to replace some foreign words in the German language by words of native origin, and there have been movements to replace later accretions in English by words derived from Old English forms. In the long run, such attempts never succeed in preventing or reversing change; at best they preserve collaterally supposedly purer forms and styles for certain purposes and in certain contexts.

Attempts to prevent or reverse linguistic change

With the picture painted above of the tendency for languages to fragment first into dialects and then into separate languages, it might be thought that dialects are relatively late in appearance in the history of a language family. This impression is reinforced by the fact that most nonstandard dialects are unrepresented as such in writing, and so comparatively little is known about dialectal differences within most languages as one goes back in time. In this respect the very detailed knowledge of the Ancient Greek dialect situation is quite untypical.

In fact, dialect divisions must have been a feature of linguistic communities as early as there is any knowledge of them. Dialect splitting is fostered by isolation and loss of contact between groups within a speech community, and the sparse populations of earlier days, often nomadic and spread over large areas relative to their numbers, will have encouraged this process. It is simply the case that all but literate dialects have been lost in the past, and an artificial homogeneity is attributed to most ancient languages and to the so-called reconstructed parent languages of families.

Present-day conditions tend toward the amalgamation of dialects and the disappearance of those spoken by relatively few people. Urbanization, mass travel, universal education, broadcasting, ease of communication, and social mobility all foster rather large regional and social dialects, with special occupational types of language within them, in place of the small, strictly localized dialects of earlier times. This is one reason for the urgency with which dialect studies are being pursued in many Western industrialized countries, such as England and parts of the United States. If work is not done soon, many dialects may perish unrecorded (see also the section *Dialects* above).

For the same reasons, dialect divisions that earlier would have widened into distinct languages are now unlikely to do so. One may compare the emergence of the separate Romance languages from once unitary Latin with the splitting of South American Spanish and Portuguese into different dialects of these two languages. These dialectal divisions are not now expected to widen beyond the range of intercomprehensibility. These same conditions, together with the spread of literacy, are leading to the extinction of languages spoken by relatively small communities. Such is the fate of most of the North American Indian languages, and Irish, Welsh, and Scots Gaelic may ultimately survive only as learned second languages, preserved as cultural focuses for their communities. But in situations like this, both past and present, the intervening period of extensive bilingualism and the concomitant use of two languages has its effect on the changes taking place in the dominant language, which is influenced by the phonetic and grammatical composition of the speakers' former language.

Extinction of languages

LANGUAGE TYPOLOGY

Language families, as conceived in the historical study of languages, should not be confused with the quite separate classifications of languages by reference to their sharing certain predominant features of grammatical structure. Such classifications give rise to what are called typological classes.

In fulfilling the requirements of open-ended creativity imposed on language by human beings, grammatical structure has things in common in all known languages, particularly at the deeper levels of grammar. All known languages have words or wordlike elements combined in accordance with rules into sentences; all known languages distinguish in some way nounlike and verblike sentence components; and all known languages have the means of embedding or subordinating one sentence within another as an included clause (*e.g.*, "the sun set" and "we returned home": "When the sun set we returned home"; "Joan was playing tennis" and "Joan twisted her ankle": "Joan, who was playing tennis, twisted her ankle," or "while she was playing tennis, Joan twisted her ankle"). Descriptive analyses of all the languages of the world have not yet been prepared, and, of course, there is information about only a minute number of those that are no longer spoken—namely, those few that were written. But there is enough known to make the assertion of such universal features as have been given with fair confidence. These are often referred to as language universals; their nature and extent is the subject of current discussion and research.

Within these very general guidelines, however, languages exhibit various types of structure. This can most readily be seen by comparing the relations between the forms of words and their syntactic functions in different languages. Such a comparison is the basis of three broad types of language that have been distinguished since the beginning of the 19th century. They are, in fact, more like characteristics than types, in that most languages contain traces of all three, in different proportions.

Broad types of language structure

Classical Chinese made little or no use of word-form variation, such as is found, for example, in Latin, for grammatical purposes. Sentence structure was expressed by word order, word grouping, and the use of specific grammatical words, or particles. Such languages have been called isolating or analytic. Modern Chinese languages are much less analytic than is often believed; probably, Vietnamese is the most fully representative of this type today. Some languages string together, or agglutinate, successive bits, each with a specific grammatical function, into the body of single words. Turkish is a typical agglutinative language: compare Turkish *evleri* "houses" (accusative case), in which *ev* is the root meaning "house," *-ler* marks plurality, and *-i* is the sign for accusative, with Latin *domūs*, in which *-ūs* combines the representation of accusative and plural without the possibility of assigning either category separately to one part of the word ending. Latin is in this respect an inflectional, or fusional, language. In a more extreme example, Latin *ī* "go!" cumulatively represents in one fused form the verb meaning "go," active voice, imperative mood, second person, and singular number, each a grammatically distinct category.

English, like many other languages, is representative of all three types. In its use of word order alone to distinguish grammatical differences ("the dog chased the cat"; "the cat chased the dog") it resembles Classical Chinese rather than Latin. In a word form such as "manliness," in which each bit can be assigned a grammatical function ("man" the basic noun, *-li-* the adjective formative, and *-ness* the abstract noun formative), it makes use of agglutination, whereas plurals such as "men" and "geese" and past tenses such as "came" and "ran" fuse distinct grammatical categories into a word form in which only arbitrarily can one allot some sound segments, or letters, to one and some to the other.

Assigning languages to different types in this way involves a delicate procedure of balancing one part of the grammar against another and deciding which type of structure predominates and how well the other types are represented. Languages predominantly of each of the types are found in communities at with all levels of civilization and with all types of culture.

Change in grammatical structures

In the course of transmission, grammatical structures change, just as do pronunciation and meanings, and in time the cumulative effect may be the transference of a language from one overall type to another, although it remains descended from the earlier language and therefore is just as much part of the same historical family. Latin is very different typologically from French in its grammatical structure, but French is nevertheless the form that Latin took in France in the course of time. In the matter of the grammatical relevance of word order, the absence of case inflections in nouns, and the use of verbal auxiliaries instead of single word tense forms, French is more like English, a distant cousin within the Indo-European family, than it is like Latin, its immediate progenitor (compare French *j'ai donné*, English "I have given," Latin *dedi*). The two sorts of language classification, historical and typological, serve different purposes and are differently based. Language families group languages together on the basis of descent; *i.e.*, unbroken transmission from an earlier common parent language. The evidence is, in the main, systematic correspondences among the shapes of words of similar meanings (*e.g.*, Greek *patēr*, Latin *pater*, French *père*, German *Vater*, English "father"). Languages are put into typological classes, with the reservation already mentioned, on the basis of certain overall similarities of structure irrespective of historical relations. Though these two classifications may coincide with some languages, as is the case to a great extent in the Bantu family, they do so only contingently; being based on different data and oriented differently, they do not logically or necessarily imply each other.

In a way, these two systems of classification involve the two most important aspects in which languages must be seen for them to be properly understood: as products of a continuous historical process and also as self-sufficient systems of communication in any one period. Both as a component of cultural history and as a central part of culture itself, language is able to reveal, more than any other human activity and achievement, what is involved in mankind's distinctive humanity. (Ro.H.R.)

BIBLIOGRAPHY. A bibliography for LANGUAGE is likely to overlap at least partially with the bibliography for LINGUISTICS. This bibliography draws attention to some books that may usefully be consulted without prior specialist knowledge, and that develop in further detail the major topics introduced in the article.

EDWARD SAPIR, *Language* (1921), one of the most attractive books on language ever written; LEONARD BLOOMFIELD, *Language* (1933), longer and more technical than Sapir's book, but a classic work that remains unmatched in its breadth of coverage, though Bloomfield's treatment of semantics is now felt to be somewhat dated through his strict adherence to behaviourist principles; JEAN F. WALLWORK, *Language and Linguistics* (1969), a short, simple, and modern introduction to the study of language, and *Language and People* (1978), a brief discussion of the social dimensions of language; HERBERT H. and EVE V. CLARK, *Psychology and Language* (1977), an introduction to theories of language acquisition; DWIGHT L. BOLINGER, *Aspects of Language* (1968), very useful wide-ranging survey of current approaches to the subject; DAVID ABERCROMBIE, *Elements of General Phonetics* (1967), and PETER N. LADEFOGED, *A Course in Phonetics* (1975), excellent introductions to the linguistic study of human speech; JOHN LYONS, *Introduction to Theoretical Linguistics* (1968), deals with linguistics rather than with language, but is an important textbook that introduces the reader to some of the most significant developments in the theory of grammar and semantics, as does the same author's later work *Language and Linguistics: An Introduction*; N. MINNIS (ed.), *Linguistics at Large* (1971), a collection of papers on different aspects of language, several of which treat at greater length some of the topics mentioned in the article; ROBERT H. ROBINS, *A Short History of Linguistics*, 2nd ed. (1979), a brief historical account of the study of language from antiquity to the present day. DELL HYMES (ed.), *Language in Culture and Society: A Reader in Linguistics and Anthropology* (1964), contains a number of useful articles on the relations between language and man's life in society. ROY HARRIS, *The Language Myth* (1981), explores the relations between language and thought; GEORGE A. MILLER, *Language and Speech* (1981), tries to explain language from the point of view of biology; ERIC WANNER and LILA R. GLEITMAN (eds.), *Language Acquisition: The State of the Art* (1982), researches how children acquire language; HANS AARSLEFF, *From Locke to Saussure: Essays on the Study of Language and Intellectual History* (1982), challenges established language theories; DAVID LIGHTFOOT, *The Language Lottery: Toward a Biology of Grammars* (1982), examines the place of language in the system of human cognition and perception; JEREMY CAMPBELL, *Grammatical Man: Information, Entropy, Language, and Life* (1982), addresses language and information theory; DEREK BICKERTON, *Roots of Language* (1981), examines origins of languages; and GRAHAM D. MARTIN, *The Architecture of Experience: A Discussion of the Role of Language and Literature in the Construction of the World* (1981), is a special study.

Languages of the World

Languages may be classified either genetically or typologically. A genetic classification assumes that certain languages are related in that they have evolved from a common ancestral language. This form of classification employs ancient records (such as those for Latin) as well as hypothetical reconstructions of the earlier forms of languages, called protolanguages. Because information on the genetic affiliations of languages is sufficiently extensive, world surveys of languages are necessarily oriented in that way—sometimes exclusively so and sometimes in conjunction with typological classifications. Typological classification is based on similarities in language structure. Individual frames of reference in language typology are not known well enough to permit a worldwide typological classification.

Before the conclusive demonstration that unwritten languages could be classified genetically, they were often relegated to a typological classification, which at one time was denigrated by scholars. Since 1917, however, the prestige of some kinds of typology has risen—in particular, that of grammatical typology. The best-known typological frame of reference represents the grammar of a language, either as a whole or as a subsystem. Once a genetic classification has been established, typological classification may be superimposed on it in order to show change of language type—as from a predominantly inflectional language (such as Proto-Germanic) to a predominantly isolating one (such as modern English)—or to show features that are shared by languages in neighbouring branches in the same family (*e.g.*, Celtic and Germanic in Indo-European). The ultimate grammatical typology is that which treats subsystems that are, in some sense, universal to all human languages.

Lexical typologies, based on similarities in vocabulary structure, have been used in cognitive anthropology and psycholinguistics (*e.g.*, perception of colours and use of colour terms). The sociolinguistic frame of reference in typology provides classifications for varieties of language in terms of their functions and their ways of identifying social groups and cultural spaces; in addition, it brings order and integration to problems concerning national standards that are faced by new nations that have many nonstandard and unwritten languages as well as languages that make use of writing.

A few points of terminology should be explained before further discussion of the world's languages is afforded. *Language family* is the label often used for a conservative genetic classification, one that can be attested only when an abundance of cognates (related words) is available. *Phylum* is the label for a liberal genetic classification that is attested with fewer cognates; it encompasses language families. Although a given phylum will have greater extension than any of the families included in it, only fragments of phonology will be reconstructible in the protolanguage. In actual linguistic usage, however, the term *family* is often employed to refer to a group that is technically a phylum—*e.g.*, the Afro-Asiatic (Hamito-Semitic) family, the Sino-Tibetan family.

The label *language isolate* is used for a language that is the only representative of a language family, as Basque or the extinct Sumerian language; the presumptive but unknown sister languages of isolates are dead and unrecorded. A language isolate may be classified, along with normal language families, under the rubric of an extensive phylum (*e.g.*, Korean is sometimes classified as a member of a hypothetical Ural-Altaic phylum) or left wholly unclassified (*e.g.*, the Ainu language of Japan). The label *pidgin-creole* is used for a language that has had so much vocabulary change that cognates for reconstructing the protolanguage from which it descended cannot be found. A pidgin is a contact language used for communication between groups having different native languages. When a pidgin becomes the native language of a community it is customarily called a creole.

This article begins with a survey of world languages based on geographic regions of unequal size: huge and sprawling areas for the peripheral regions of Africa, Oceania, and the Americas but relatively compact areas for the focal regions within Eurasia. Nine regions—six in Eurasia, in all of which writing and standard languages are widespread—constitute a convenient basis for comparison and contrast. The larger part of the article consists of more detailed examinations of the languages of the world arranged by genetic affinities.

Facets of the subject of language and human communication are treated in a variety of *Macropædia* articles. For a full account of the theory and methods of linguistic science, see the article LINGUISTICS. For information on such subjects as the characteristics of language, language variants (slang, jargon), speech production, and the acquisition of language, see the article LANGUAGE. For a full account of phonetics and the pathology of speech, see the article SPEECH. For information on written languages and writing systems, see the article WRITING.

For coverage of related topics in the *Macropædia* and the *Micropædia*, see the *Propædia*, section 514, and the *Index*.

The article is divided into the following sections:

INTRODUCTORY SURVEY

LANGUAGES OF EUROPE

The great majority of the languages spoken in Europe are of Indo-European and Uralic (especially Finno-Ugric) affiliation. In terms of numbers of speakers, however, the people in Europe who speak the languages of these families are now fewer than those in non-European countries who also speak such languages. For example, Latin America (rather than Europe) is now the chief locus of the Spanish language.

An unusually small degree of genetic diversity is found among European languages: there are fewer language families in Europe than in any other continent-size region of the world. In addition, literary traditions that have resulted in the preservation of earlier forms of present-day languages are found to a high degree among these languages. Every European language with a writing tradition has developed at least one standard that is recognized nationally, and the national standard often coexists with recognized regional standards.

A few European languages are used internationally, as lingua francas, but there is a low degree of pidgin and creole usage in Europe today.

Typological classifications have been superimposed on genetic classifications of European languages in particular. For example, the Italic branch of Indo-European languages may be grouped with the Greek, Celtic, and Germanic branches on the basis of certain structural features, as can Armenian with Greek and the Indo-Iranian languages, and so forth.

Indo-European languages. The languages of seven of the nine extant branches of the Indo-European language family are spoken in Europe. Variability in estimates of the number of particular languages reflects variation in the criteria used (*e.g.*, mutual intelligibility between neighbouring dialects, known common or separate history versus sociocultural factors such as separate literary traditions or status as national languages of politically independent units) as well as in the time period for which the criteria are applied. Thus, for example, it is possible to say on linguistic grounds that there are nine extant languages in the Romance subgroup of the Italic branch: Portuguese, Spanish, Catalan, French, Romansh, Ladin, Friulian, Italian, and Romanian. In applying the criterion of separate literary tradition, the list would be expanded by the addition of Provençal and Sardinian. To apply the criterion of status as a national language would reduce the list because Ladin, Friulian, Provençal, and Sardinian are not national languages; but the picture is complicated by the fact that Sardinia was once politically independent, and Andorra, in which Catalan is spoken, has not always been independent.

Criteria for determining the number of languages

Similar difficulties in counting separate languages exist for all the branches in which several languages are spoken. In terms of areas of high mutual intelligibility (which do not entirely reflect historical development), there are only five modern Germanic languages: English, Frisian, Netherlandic-German (including Afrikaans and Yiddish), Insular Scandinavian, and Continental Scandinavian. If literary tradition and national criteria are considered, the number is increased by the division of Netherlandic-German into Standard High German, Low German, Dutch-Flemish, Afrikaans, Luxemburgian, and Yiddish; the division of Insular Scandinavian into Icelandic and Faroese (Faeroese); and the division of Continental Scandinavian into Norwegian (which is further subdivided into New Norwegian [or Nynorsk] and Dano-Norwegian [or Bokmål]), Danish, and Swedish.

For the Slavic languages there are 13 literary standards, but between the nuclei formed by these norms there are scarcely any linguistic boundaries, because transitional dialects connect adjacent areas. In terms of intelligibility and, to some extent, in terms of shared features, the Slavic literary norms can be grouped into three zones: East Slavic (including Russian, Belarusian, and Ukrainian), West Slavic (including Polish, Kashubian, Lower, or Low Sorbian, Upper, or High Sorbian, Czech, and Slovak), and South Slavic (including Slovene, Serbo-Croatian, Macedonian, and Bulgarian).

Language boundaries are more clear-cut for the modern living languages in the remaining European branches of Indo-European: Celtic (the physical separation of the speakers of the languages contributes to the separate identification of Welsh, Breton, Irish, and Scottish Gaelic), Baltic (the literary and political separation coincide with the separation of Lithuanian and Latvian), Greek (the separate historical development and lack of mutual intelligibility separate Modern Greek from Tsakonian), and Albanian (the political unity of Albania contributes to the single-language identification of Gheg and Tosk, its two divergent dialects).

Dialects of two languages in the Indo-Iranian branch of Indo-European also are or were spoken in Europe: the Jassic dialect of Ossetic, an Iranian language, formerly spoken in Hungary; and the European dialects of Romany, which was spread by Gypsies throughout Europe and into America. It may be, however, that only in Wales, Finland, and the Balkans does Romany still serve as a native language, though Welsh and Finnish Romany have few speakers left.

Extinct Indo-European languages

A number of earlier Indo-European languages that died out without descendants are known from written records and comments by contemporaries; these include several in the Italic branch (such as Oscan, Umbrian, Faliscan, Venetic), at least four in the East Germanic group (including Burgundian, Ostrogothic, Visigothic, and Vandalic), and three in the Celtic branch (Gaulish, Cornish, and the recently extinct Manx). The classification as Indo-European or non-Indo-European for many other extinct languages (such as Pictish, spoken in what is now northern Scotland) remains uncertain because of the scarcity of data.

For more information on the Indo-European languages of Europe, see below *Greek language; Italic languages; Romance languages; Germanic languages; English language; Celtic languages; Baltic languages; Slavic languages; Albanian language.*

Uralic languages. In addition to the Indo-European languages, all but two of the languages of the Finno-Ugric branch of the Uralic family also are spoken in Europe. As in the case of Indo-European languages, variation exists in the enumeration of separate languages. For example, several varieties of Sami (Lapp) are mutually unintelligible but often are classified as dialects of a single language. The various types also have been classified according to geographic areas or national boundaries (Norwegian, Swedish, Finnish, and Russian Sami). In Baltic-Finnic, the Finno-Ugric subgroup most closely related to Sami, the Finnish, Karelian, Veps, Ingrian, Estonian, Livonian, and Votic languages often are linked by transitional dialects between the central areas of a given pair. The other Finnic languages, Mari and Mordvin, and all three languages of the Permic subgroup, Udmurt (formerly Votyak), Permyak, and Komi, are spoken much farther to the east, in the central area of eastern European Russia.

The Samoyedic branch of Uralic is represented by the Nenets, who live in northernmost Russia from the mouth of the Northern Dvina River eastward into North Asia. The remaining Uralic language of Europe, which belongs to the Ugric subgroup, is Hungarian. See below *Uralic languages.*

Other languages. *Maltese.* Maltese, which is spoken in Malta, is a Semitic language descended from a dialect of Arabic. It was so long isolated from other dialects of Arabic and so heavily influenced by Italian that the resultant loss of mutual intelligibility with other Arabic speakers justifies its usual classification as a separate Semitic language.

Basque. Basque, spoken in the Pyrenees in Spain and France, is the only other living language of western Europe that does not belong to the Indo-European family. Numerous attempts to link Basque genetically to other languages have been inconclusive. See below *Language isolates: Basque language.*

Turkic languages of eastern Europe

Turkic languages. In addition to Turkish, which is spoken by a number of people in Bulgaria and elsewhere in the Balkans, several languages of the Turkic language group (classified as a subfamily of the Altaic language family) are spoken entirely in eastern Europe. Chuvash, the most divergent Turkic language, is found mainly in Chuvashia in Russia; Tatar in Tatarstan and adjacent areas and in Romania and Bulgaria; Bashkir in Bashkortostan; Gagauz in Ukraine and Moldova and in the Balkans; and Karaim in southern Ukraine and Lithuania. Most or all of the speakers of Crimean Turkish were removed to the Uzbek S.S.R. (now Uzbekistan) after World War II, although since 1989 a number have returned to the Crimean Peninsula. See below *Altaic languages.*

Extinct languages. The existence of a number of long-extinct non-Indo-European languages of Europe is known through the records of the Greeks and Romans and also through the preservation of varying amounts of written records. The most extensive records are those in the still undeciphered Etruscan, which is known to have been spoken in Italy from the 8th century BC to the 4th century AD (see below *Language isolates: Etruscan language*). Several languages were spoken in the Iberian Peninsula, of which Iberian (preserved in a few inscriptions and many coins) was spoken along the Ebro River and at one time as far east as the Rhône River.

LANGUAGES OF SOUTH ASIA

The genetic classification of the languages of India, Sri Lanka, Bangladesh, Pakistan, Nepal, and Bhutan includes two subgroups—Indo-Aryan (also called Indic) and Iranian—of a single branch of Indo-European (called Indo-Iranian), some indigenous language families (such as Dravidian), a few language isolates (such as Burushaski), and some Sino-Tibetan languages.

Indo-Iranian languages. Except for Romany and the few Dardic languages spoken in Afghanistan, all the languages of the Indo-Aryan subgroup of the Indo-Iranian branch of Indo-European are spoken in South Asia. It is difficult to identify language boundaries in the Indo-Aryan group because, between any pair of literary standards, "transitional" dialects grade into one another, with no clear-cut language barriers. The problem is further complicated by the enormous dialect differentiation in most of the Indo-Aryan languages. In terms of lack of mutual intelligibility between literary standards, there are more than 20 Indo-Aryan languages. Although Sanskrit is a classical Indo-Aryan language, preserved in writing, it also enters so deeply into the vocabulary of present-day languages as to become, in some cases, the salient mark differentiating two dialects of one language. Thus, Hindī of India differs linguistically from Urdū of Pakistan chiefly in that the former may be heavily Sanskritized in vocabulary and the latter not.

Ethnolinguistic loyalties also may increase the number of languages distinguished—*e.g.,* the separate recognition of Bengali and Assamese. Most of the Indo-Aryan languages are spoken by many millions of speakers—*e.g.,* Bengali, Assamese, Hindī, Marāṭhī, Maithilī, Maghī, Bhojpurī, Gujarātī, Oṛiyā, Sinhalese (in Sri Lanka), Sindhī, and Nepālī. There are also large numbers of speakers of Indo-Aryan languages (especially West Hindī) in South Africa, in various parts of the South Pacific, and in South America (particularly Suriname and Guyana).

The Dardic subgroup of Indo-Iranian

The languages of the Dardic subgroup differ sufficiently from the other Indo-Aryan languages as to be sometimes classified as Iranian rather than Indo-Aryan or as a separate subbranch coordinate with the Indo-Aryan and Iranian subbranches. Kashmirī, spoken in Jammu and Kashmir, is the only Dardic language with a literary tradition. Shīnā also is spoken in Jammu and Kashmir; other Dardic languages, which are spoken mostly in Pakistan, have relatively few speakers.

Some speakers of at least four Iranian languages also are found in South Asia, including Pashto speakers in Pakistan and Balochi (Baluchi) speakers in both Pakistan and India (see below *Indo-European languages: Indo-Iranian languages*).

Dravidian languages. Although the greatest concentrations of Dravidian speakers are in southern India, the more than 20 languages of this family are widespread in India, and one language, Brāhūī, is isolated in Pakistan, separated from its nearest sister language by 800 miles. Four Dravidian languages have long literary traditions and are spoken by many millions: Telugu, Tamil, Malayālam, and Kannaḍa. Tamil speakers also are found in Sri Lanka, Malaysia, Indonesia, Myanmar (Burma), Vietnam, and

South Africa and in scattered island and coastal areas around the world. Among other, less widespread Dravidian tongues of India are Gondī, Tulu, Kurukh, and Kui. No convincing remote relationships between the Dravidian family and other families have been proposed (see below *Dravidian languages*).

Austroasiatic languages. The 16 or so Muṇḍā languages are all spoken in India. Most scholars classify them as a language family within the Austroasiatic stock. Santhālī is the Muṇḍā language with the greatest number of speakers (a few million); Muṇḍārī, Ho, Sora, Khaṛiā, and Korku have significantly fewer speakers. Some scholars include Nahali, spoken by a few hundred people in southwestern Madhya Pradesh, among the Muṇḍā languages. Khāsī, spoken by a majority of the population of the Indian state of Meghālaya, is a member of the Mon-Khmer language family. See below *Austroasiatic languages.*

Sino-Tibetan languages. Speakers of languages of most of the branches of the Sino-Tibetan language family are to be found in South Asia. All the languages of the Bodo (Bodo-Gāro) branch of the Bodo-Nāga-Kachin language group are spoken in Assam. Nāga languages are spoken in scattered locations from eastern Nepal into Myanmar. The Kachin (Ching-p'o) languages are centred in northern Myanmar, but some dialects are spoken in Assam, where there are also some speakers of Kuki (Kuki-Chin) languages and of Burmese (Burmese-Lolo) languages. Dialects of the various divisions of the Tibetan language are distributed from Kashmir to Bhutan and southward into India (*e.g.*, Bāltī, Sherpa, Lhoke, Spiti). Speakers of close to 50 Gyarung-Mishmi (or Himalayan) languages are found in northeastern India, with their greatest concentration in Nepal. (See below *Sino-Tibetan languages*).

Other languages. *Tai languages.* Some speakers of Khamti, a Tai language, live in Assam, where Ahom, another Tai language, is still used as a ceremonial language in religious rituals but is no longer spoken. See below *Tai languages.*

Use of English. In all parts of postcolonial South Asia, including Sri Lanka, some people know English; these speakers, although relatively few in number, are the people most likely to travel to a state in which a South Asian language unknown to them is spoken. Hence, English is de facto the current interstate and international language of South Asia, although many Indians would prefer to adopt another language, such as Hindī or a Dravidian language, as the national language.

English as lingua franca

Burushaski. Burushaski, spoken by some 50,000 people in far northeastern Pakistan, is without even remote known relatives.

LANGUAGES OF NORTH ASIA

The languages of North Asia are those spoken from the Arctic Ocean to South Asia and China and from the Caspian Sea and Ural Mountains in the west to the Pacific Ocean in the east. In genetic classification most languages of North Asia belong to the Uralic family, to one of the three families of the Altaic stock (Turkic, Mongolian, and Manchu-Tungus), or to Indo-European. The genetic affiliations of the Paleo-Siberian languages, spoken exclusively in this region, are uncertain at present. Scholars have hypothesized that some of the languages once may have been American Indian languages whose prehistoric speakers backtracked from the New World into North Asia. That preindustrial peoples traversed the Arctic waters is demonstrated by the presence of Eskimos on both the Russian and Alaskan shores of the Bering Strait. Some scholars have claimed that all languages indigenous to North Asia, except the Paleo-Siberian ones and the recently intrusive Russian language, are genetically related in a Ural-Altaic phylum. This liberal classification, however, is questioned by many scholars.

Whether or not Altaic and Uralic are related to each other, there is no doubt that languages of both stocks share many typological features, such as vowel harmony, agglutination (a type of word formation in which word elements are added together but still retain a separate, definite meaning), and a restriction against combining a plural noun with a quantifier (as though, in English, the plural noun "girls" had to appear as a singular noun in the phrase "five girl," rather than "five girls").

Uralic languages. Languages from three branches of Uralic are spoken in North Asia; their speakers are few in number. The Yukaghir family, which may be an offshoot of Early Uralic, includes one living language, Yukaghir (spoken by a few hundred people south of the Arctic Circle on tributaries of the Kolyma River and in the tundra between the Indigirka and Alazeya rivers). An extinct Yukaghir language, Chuvan (or Chuvantsy), was spoken until the 20th century on the Anadyr River. The two Ob-Ugric languages—Mansi (also called Vogul) and Khanty (also called Ostyak)—are spoken on the Ob River and its southwestern tributaries. All the languages of the Samoyedic branch are spoken in North Asia: Nenets, speakers of which are scattered from the mouth of the Yenisey River westward to the mouth of the Northern Dvina; Enets, also centred on the Yenisey; Nganasan, spoken on the Taymyr Peninsula in Siberia; and Selkup, spoken in a region lying south of that of the Enets speakers, between the Taz and Tym rivers. Another southern Samoyedic language, Kamas (Sayan), was functionally extinct by 1987 (see below *Uralic languages*).

Altaic languages. *Turkic languages.* The Turkic languages are remarkable for their lack of diversity in spite of their wide occurrence in all the Eurasian regions except South Asia and Southeast Asia. Several of the numerous Turkic languages might be considered as a single language if it were not for the fact that mutual intelligibility between groups is impaired by differential borrowing from the various unrelated languages encountered in different regions. Thus, a Turkish speaker from Turkey might understand or largely understand an Uzbek speaker with more ease than the Uzbek speaker would understand Turkish, which has many loanwords from Persian and Arabic. Educated speakers of Turkic languages are able to read written materials in other Turkic languages after some adjustment to their varying spelling conventions and sound correspondences. Because such differences are identified with different Turkic ethnic groups, it is customary to identify the larger of these ethnic groups as speaking different Turkic languages, although some degree of intelligibility exists between them, as it does between the Uzbek, Bashkir, and Tatar languages. In addition, some languages have dialects that are transitional between two recognized language groups; *e.g.*, some dialects of Karakalpak are said to be transitional to Turkmen, and others are said to be transitional to Uzbek.

Relative uniformity of Turkic languages

In North Asia, Turkic languages are distributed from the southern extension of North Asia northeastward through central Siberia and include Turkmen in Turkmenistan, Iran, and Afghanistan; Uzbek in the Central Asian countries, primarily Uzbekistan, and in Afghanistan; Kyrgyz in Kyrgyzstan and in neighbouring areas from Afghanistan to China; Karakalpak in Karakalpakstan republic in Uzbekistan; and Kazakh in Kazakhstan. Six or seven Turkic groups immediately north of western Mongolia are much smaller in terms of both numbers of speakers and the region they inhabit; in northern Siberia the Yakut extend from Sakha republic (Yakutia) west to Taymyr autonomous *okrug* (district), where the Dolgan people speak a Yakut dialect.

Mongolian languages. The Mongolian languages are dispersed throughout Central Asia from Afghanistan to Manchuria, occupying large parts of North and East Asia. The problem of recognizing language boundaries (*i.e.*, of distinguishing separate languages) in the Mongolian family is complicated by the fact that differences between dialects are exaggerated in areas where Mongolian speakers have borrowed features of different unrelated languages but are minimized in areas where one dialect is spoken as a lingua franca throughout an extensive region. Among the Mongolian languages are Mogol, spoken in Afghanistan, where it has been influenced by Iranian and Turkic languages; Monguor, spoken in Kansu province of China and in Tibet, with noticeable effects of both Tibetan and Chinese in the language; and Daghur, spoken mainly in Inner Mongolia and heavily influenced by Tungus languages. Additional languages include Ordos in Inner Mongolia, Kharachin in China, Oyrat in the Sino-Russian border area from Kyrgyzstan to the Altai Mountains, and Buryat from Buryatia into Inner Mongolia. Some degree of mutual intelligibility exists between some of these, but this may be in part the result of the lingua franca use of Khalkha, the official language of Mongolia.

Manchu-Tungus languages. Speakers of the Manchu-Tungus languages are scattered from central interior Siberia to the shores of the seas of Japan and Okhotsk, including the Kamchatka Peninsula and Sakhalin Island. Those not near the coast live generally along the banks of the major rivers—the Yenisey, Tunguska, Khatanga, Lena, Amur, and Sungari. Detailed information on most of the Manchu-Tungus languages is scanty, and language names usually coincide with politico-cultural groups, rather than being based on a comparison of linguistic features or knowledge of mutual intelligibility. Borrowing that resulted from contact with speakers of Samoyedic (Uralic) languages to the west and northwest, Mongolian languages and Sinitic (Chinese) languages to the south, and the various Paleo-Siberian languages to the north and east has further complicated the subclassification of the Manchu-Tungus languages by increasing the superficial differences among them.

Difficulties in Manchu-Tungus classification

Most speakers of Manchu-Tungus languages are bilingual in the official language of their country, and many are replacing their native languages with Russian or Chinese. After the Manchu in China, the next best known and numerous of the Manchu-Tungus peoples are the Evenk. Other groups include the Even (or Lamut), Nanai (or Gold), and other relatively small tribes. For more information on the Turkic, Mongolian, and Manchu-Tungus languages, see below *Altaic languages.*

Paleo-Siberian languages. Most of the peoples whose languages are grouped together under the catchall category of Paleo-Siberian live in northeasternmost Siberia in the area between the East Siberian Sea and the Sea of Okhotsk, including the Kamchatka Peninsula, along the coast of the Sea of Okhotsk as far south as the Amur River, and on Sakhalin Island; peoples of another Paleo-Siberian group live far to the west along the middle and upper Yenisey River. The Paleo-Siberian languages form three groups that are not only not related to each other but also have not been demonstrated to be related to any other genetic groups. (Another group sometimes classified as Paleo-Siberian, the Yukaghir, is now considered by some linguists to be a member of the Uralic language family, perhaps an offshoot of Early Uralic.)

The northernmost and most widespread of these linguistic groups and the only one that includes more than one living language is the Luorawetlan family, which consists of Chukchi, Itelmen (Kamchadal), and Koryak. Most scholars now classify Kerek and Aliutor, once considered to be dialects of Koryak, as independent languages.

Nivkh (Gilyak), spoken on Sakhalin Island and in the coastal and inland Amur River country of the mainland, has no known linguistic relatives. Ket (or Yenisey-Ostyak) is the only language of the Yeniseian or Yenisey-Ostyak family that is still spoken. Ket speakers live along the upper and middle Yenisey River, as did the speakers of its sister languages, Kott (Cottian-Manu), which became extinct in the 19th century, and Assan (Asan) and Arin, both of which became extinct in the 18th century (see below *Paleo-Siberian languages*).

Other languages. Indo-European languages, like Russian (a Slavic language), were introduced to North Asia only relatively recently. They include the Iranian languages in the southwestern extension of North Asia (Tajik, or Western Fārsī [Persian], in Tajikistan and Balochi in Turkmenistan) and the long-extinct Tocharian, which penetrated into Central Asia as far as Chinese Turkistan (see below *Indo-European languages: Indo-Iranian languages* and *Tocharian language*).

Writing and literacy in North Asia. The earliest stimulus toward writing in North Asia was from China. The latest stimulus, from Soviet Russia, brought literacy to those Altaic peoples whose languages were unwritten in tsarist times. In contrast to the written Altaic languages, the Paleo-Siberian languages in general remain unwritten. Soviet educational policy encouraged the use of native languages for education and for teaching writing. The standard form of writing Tajik, for example, is in the Cyrillic alphabet, and knowledge of this alphabet facilitated later learning of Russian, which is the modern lingua franca of North Asia. In the Mongol empire of the 13th century, Turkic languages were used as languages of administration across North Asia from the Caspian Sea to Manchuria and, initially, in adjacent Eurasian regions conquered by Genghis Khan.

Use of the Cyrillic alphabet

LANGUAGES OF SOUTHWEST ASIA

Languages spoken today in the area from Iran westward to the Mediterranean (in Iran, Iraq, Saudi Arabia, Jordan, Syria, Lebanon, and Israel) are Semitic, Indo-European, or Turkic. The languages in the two marginal subareas of Southwest Asia (in Afghanistan and in Turkey and the Caucasus Mountains between the Black and Caspian seas) far exceed the languages of Europe in genetic diversity.

At one time Sumerian, which is preserved in written form, was spoken as the first language of civilization in the ancient Middle East; this language was neither Semitic nor Indo-European (see also *Language isolates: Sumerian language*). Early literary traditions and literacy for the elite began in this central area of Southwest Asia and extended from the Sumerian, Old Persian, and Akkadian literatures to Asia Minor (Hittite) in the north and to the Nile River (Egyptian) in northeastern Africa. Akkadian and Persian seem to have been the first two languages put to wide international use.

Indo-Iranian languages. Almost all of the score of living languages of the Iranian subgroup of the Indo-Iranian branch of Indo-European are spoken in Southwest Asia, occasionally extending into neighbouring regions. Persian has three separate literary standards that are not confined to the countries in which they centre (Iran, Afghanistan, and Tajikistan). More than half of the speakers of Pashto live in Afghanistan and the rest in South Asia. Kurdish is spoken in an area extending southward from southern Armenia into Turkey, Syria, Iran, and Iraq. Perhaps three-fifths of the speakers of Balochi live in Iran and southern Afghanistan. Several other Iranian languages (or dialects) have many fewer speakers; these include Lurī and Bakhtyārī, spoken only in Iran, and Munjī and Shughnī, spoken largely in Afghanistan, with only a few of their speakers in Pakistan or Tajikistan. One Iranian language, Yaghnābī, is spoken only in Tajikistan. Three Iranian languages are spoken almost entirely in the Caucasus: Tat, Talysh (with some speakers in Iran), and Ossetic.

The half a dozen Nūristānī languages spoken in Afghanistan and part of Pakistan, sometimes classified as members of the Dardic subgroup of Indo-Aryan, more recently have been classified by some scholars as constituting a separate branch of Indo-Iranian. In addition, some Laṇḍā (Indo-Aryan) speakers also live in Afghanistan. Two very divergent dialects of another Indo-Aryan language, Romany, are spoken in Southwest Asia—Armenian Romany and Asiatic Romany (the dialect of the Palestinian Gypsies). For more information on the Iranian and Indo-Aryan languages, see below *Indo-European languages: Indo-Iranian languages.*

Other languages. The sole language of another branch of Indo-European, Armenian, is spoken predominantly in Armenia but also in Syria, Georgia, Russia, Azerbaijan, Iran, and other parts of Southwest Asia (see below *Indo-European languages: Armenian language*).

The Armenian language

The long-extinct languages of the Anatolian branch of Indo-European, including Hittite, were once spoken in Southwest Asia (see below *Indo-European languages: Anatolian languages*).

Five Turkic languages are spoken primarily in Southwest Asia: Turkish, spoken in Turkey and surrounding countries largely to the north; Azerbaijani, spoken chiefly in Azerbaijan and Iran; and Kumyk, Karachay, and Nogay, spoken in the Caucasus. Three Turkic languages spoken predominantly in North Asia also are spoken in Central Asia: Uzbek, Turkmen, and Kyrgyz.

One language of the Mongolian family is spoken in Southwest Asia—Mogol in Afghanistan; and Brahui, a

Dravidian language, has a small fraction of its speakers in Afghanistan and Iran.

Caucasian languages. In addition to the Indo-European and Turkic languages spoken in the Caucasus, there are more than 30 languages belonging to three Caucasian language families. These may be related remotely to each other in a Caucasian phylum, in which the Northeast Caucasian family is more clearly related to the Northwest Caucasian family than the South Caucasian (Kartvelian) family is to either. Georgian, a South Caucasian language, is the most widely known of the Caucasian languages, with speakers in Georgia, Azerbaijan, and adjacent parts of Turkey and Iran; it is the only Caucasian language with a long literary tradition. Other South Caucasian languages are Laz and Svan. The Northwest Caucasian (Abkhazo-Adyghian) languages include Kabardian, Abkhaz, Abaza, Adyghian, and Ubykh (the latter now extinct). The approximately 25 languages of the Northeast Caucasian (Nakho-Dagestanian) family are spoken by people living mostly in the Dagestan republic in Russia. These languages include Chechen, Ingush, Avar, Dargin (Dargwa), Lak, Lezgian, and Tabasaran. There is some scholarly disagreement concerning the classification of the Caucasian languages (see below *Caucasian languages*).

Semitic languages. Five Semitic languages are still spoken in Southwest Asia: Arabic; Hebrew, primarily in Israel; dialects of East Aramaic, still spoken in Israel, Syria, Iran, Iraq, and Armenia; West Aramaic dialects, still spoken in Lebanon and Syria; and Modern South Arabic, spoken in southern Saudi Arabia and on nearby islands. Of the extinct Semitic languages, the best known are Phoenician, Akkadian (Babylonian and Assyrian), Syriac, Moabite, and Ugaritic (see below *Afro-Asiatic (Hamito-Semitic) languages: Semitic languages*).

LANGUAGES OF EAST ASIA

Languages in East Asia are those traditionally spoken in China, Japan, and Korea—*i.e.*, those that occupy the region between North Asia and Southeast Asia. A conservative genetic classification reflects immense genetic diversity for East Asia by claiming that Ainu, Japanese, and Korean are related neither to each other nor to any other language in East Asia and that the Chinese languages (or dialects) belong in one family, Miao-Yao (Hmong-Mien) languages in another, and Tai languages in still another. A liberal genetic classification leaves Ainu isolated, includes Korean and Japanese in the Altaic family, and classifies some or all of the other groups as Sino-Tibetan.

Three types of syntax

Three general types of syntax, which partly overlap the liberal genetic classification, can be distinguished among languages in East Asia. First, Ainu is isolated syntactically as well as genetically. The second type is shared by Korean and Japanese. All Chinese languages are strikingly alike in syntax, and this third type is approximated among some non-Chinese languages of the Sino-Tibetan family and among some languages of Southeast Asia whose genetic classification is tentatively indeterminate.

Altaic languages. Languages from three of the major families of North Asia are spoken in China. Uighur, a Turkic language, is spoken in Sinkiang and Kansu provinces of China as well as in the countries of Central Asia and in southwestern Mongolia. Another Turkic language, Kyrgyz, has some speakers in China. Manchu is the best known of the Manchu-Tungus languages and that with the longest literary tradition (dating from as early as 1599). After the Manchu established the last Chinese dynasty in 1644, their language was gradually replaced in most parts of China by Mandarin—except for formal and ceremonial occasions—but it is still spoken in scattered localities in Northeast and Northwest China.

Striking similarities in syntax have led some linguists to postulate a remote relationship between the Altaic languages and Korean and, less frequently, Japanese.

Korean, Japanese, and Ainu. Korean is spoken in Korea as well as by sizable populations in China and Japan (see below *Korean language*).

The Japanese language family includes, besides Japanese, several mutually unintelligible dialects spoken on the Ryukyu Islands by people who are bilingual in mainland Japanese. Japanese is spoken by some 125 million people in Japan and by small groups in Brazil and the United States, especially in Hawaii (see below *Japanese language*).

Ainu, the remaining language in insular East Asia for which not even a remote relationship with other languages seems likely, originally was spoken in Japan and on Sakhalin Island and the Kuril Islands. By the late 20th century it was virtually extinct, with only a few speakers in northern Japan.

Sino-Tibetan languages. *Chinese languages (dialects).* Most important in terms of numbers of speakers and their influence on the other languages in East and Southeast Asia are the Chinese languages (often called dialects). In terms of mutual intelligibility among adjacent dialects, there are several Chinese languages: Mandarin, Wu, Yüeh (Cantonese), Hsiang, Kan, Hakka, and Min (or North Min and South Min). Mandarin is the native language of more than 70 percent of the Chinese and is spoken as a second language by many of the native speakers of the other languages, both Chinese and non-Chinese, in China. It has traditionally been the language of administration.

Although speakers of two different Chinese languages may not be able to understand one another when they talk, communication between them is possible in writing; conversely, the same written message is read aloud differently by speakers of different Chinese languages. The functional advantages of Chinese writing explains its perseverance for four millennia, but these advantages are partly offset by the difficulties each generation must experience in learning the thousands of character signs that are needed for literacy. Traditionally most Chinese citizens were believed to be illiterate, but, with simplified characters and romanization, the majority of the people in China are now literate. The Chinese languages are notable for their enormous numbers of speakers, and Mandarin has the largest number of speakers of any of the world's languages (some 800 million native speakers).

Literacy in China

A remote relationship in one family (Sino-Tibetan) has been postulated for the Chinese languages and all the other non-Altaic families that have languages spoken in China. In spite of the fact that there is no doubt that all these languages bear many similarities to Chinese, current knowledge fails to reveal to what extent such similarities might be the result of borrowing rather than common origin.

Tibeto-Burman languages. The Tibetan, or Tibetic, language group includes at least two Tibetan proper languages spoken in Tibet, Nepal, and India: Central Tibetan, including Lhasa, the standard dialect of Tibet, and Western Tibetan. In addition there are many other languages in Nepal, India, and Bangladesh that are closely related to Tibetan proper.

More distantly related Tibeto-Burman languages are spoken in East Asia over the borders of Myanmar (Burma); these languages, often called Burmic, include dialects of the Burmese-Lolo subgroup (including Burmese) and the Kachin subgroup. For more information on the Chinese, Tibetan, and Burmic languages, see below *Sino-Tibetan languages*.

Tai and Miao-Yao (Hmong-Mien) languages. All the languages of the Kam-Sui language group, which is related to the Tai family, are spoken in China (in Kweichow, Hunan, and Kwangsi provinces), with some dialects extending into Southeast Asia. Speakers of Miao-Yao languages are scattered over south-central China and extend into Vietnam, Laos, and Thailand. Dialects of the Miao language include Red Miao, White Miao, Green or Blue Miao, and the more divergent Black Miao. The Yao languages are Yao (also called Man or Mien), Laka, and Punu.

LANGUAGES OF SOUTHEAST ASIA (INCLUDING AUSTRONESIAN)

Southeast Asia is generally taken to be a region that includes both a mainland subregion, south of China and east of India, and an insular subregion, which includes the insular half of Malaysia, all of Indonesia, and the Philippines. Virtually all the languages of insular Southeast Asia belong to a single language family—Austronesian (Malayo-Polynesian). Mainland Southeast

Asia, on the other hand, has various representatives from the Austroasiatic, Tai, and Sino-Tibetan language groups. Hence, genetic diversity is greater in mainland than in insular Southeast Asia. Austronesian languages extend out of Southeast Asia to the most distant culture areas in Oceania (Polynesia and Micronesia), where they are the only languages known aboriginally. One modern Austronesian language (Malagasy) is spoken on the African side of the Indian Ocean in Madagascar.

Diversity in Melanesia

Curiously enough, it is in Melanesia, between the Bismarck Archipelago and Vanuatu, that the most diverse Austronesian languages are spoken today; this provides grounds for the conjecture that the Proto-Austronesian language was spoken there millennia ago and that the daughter languages diversified as their speakers migrated throughout much of the world, with Malay and Cham backtracking eventually to mainland Southeast Asia, out of which the ancestors of Proto-Austronesian speakers must have come.

In general, the name of the country and the name of the national language are the same in both insular and mainland Southeast Asia. Thus, Pilipino (based on Tagalog) is the name of one of the national languages of the Philippines, even though Pilipino is learned as a second language by most Filipinos. The fear in all of Southeast Asia of indirect neocolonial domination motivates continued distrust of the old languages of colonialism—English, French, Dutch, Spanish—and now also of Japanese and Russian. A pidgin-creole—Neo-Melanesian, or Melanesian Pidgin English—is used as a lingua franca by speakers of Austronesian and other languages from southern Papua through Melanesia into Micronesia.

Though the languages in the mainland subregion of Southeast Asia are genetically diverse, they show widespread ranges of the same typological features—such as the use of distinctive tones and classifiers—among unrelated or only remotely related languages.

Austroasiatic languages. The Mon-Khmer language family includes more than 80 languages—more than any other family that is centred primarily or entirely in Southeast Asia. Mon-Khmer languages are spoken from Myanmar to Vietnam. In Cambodia, Khmer (Cambodian) is the official language; its speakers also are found in Thailand. Mon also is spoken in Thailand and Myanmar.

The language of mainland Southeast Asia with the greatest number of speakers is Vietnamese, spoken in Vietnam and by smaller numbers of speakers in Cambodia, Thailand, and Laos. Muong, spoken in the central highlands of northern Vietnam, is recognized as a separate, but related, language and shows far less Chinese influence.

Classified as a northern group of the Mon-Khmer family are several languages spoken in Myanmar (east of Mandalay), northwestern Thailand, northern Laos, and to a lesser extent in northern Vietnam and in southwestern China. These include languages of the Palaungic, or Palaung-Wa, branch, spoken in Myanmar, Thailand, China, and Laos; and the Khmuic branch, spoken in Laos, Thailand, and Vietnam.

Another branch of Mon-Khmer, the Aslian branch, is composed of three small groups of related languages in Malaysia. They are the North Aslian, or Semang, subbranch, spoken in the inland area of northern and central Malaysia and across the border in Thailand; the Senoic, or Sakai, subbranch, with speakers south of Kuala Lumpur on the coast and inland farther south; and the Semelaic, or South Aslian, subbranch, spoken south of the Senoic languages. Data on the Nicobarese languages, spoken on the Nicobar Islands, suggest that they form a distinct branch (Nicobarese) of the Mon-Khmer family (see below *Austroasiatic languages*).

Tai and Sino-Tibetan languages. At least a dozen languages of the Tai language family are spoken in Southeast Asia: Thai, or Siamese, in Thailand; Lao, in Thailand, Laos, and Cambodia; Yuan, in Thailand; Shan, in Myanmar; Black Tai (Tai Noir), in Laos and Vietnam; Khün and Khamti, in Myanmar; and White Tai (Tai Blanc), Tho (Tay), Nung, and Kelao (Ch'i-lao), all in Vietnam.

The Chinese in Southeast Asia

Many millions of Chinese are distributed throughout Southeast Asia; of these, more than 7 million are in Thailand, 1.7 million in Malaysia, 1 million in Vietnam, and smaller numbers in Myanmar, Cambodia, and Laos.

Of the other language groups in the Sino-Tibetan family in Southeast Asia, the Burmese-Lolo (Burmish) group has the widest distribution and the greatest number of speakers. Burmese is spoken as a second language by perhaps 90 percent of those in Myanmar who have another first or native language. The Lolo languages are spoken in Myanmar, Thailand, Laos, and Vietnam; they include Lisu, Lahu, Akha, Mung, Punoi, Pyen, and others, a few of which extend into northeastern India. Karen languages are spoken in Myanmar and Thailand and include Sgaw, Pho, Pa-o (or Taungthu), and Palaychi. Most of the languages of the Kuki-Chin (Kukish) group are spoken in Myanmar. Kachin languages also are spoken in Myanmar (see below *Sino-Tibetan languages* and *Tai languages*).

Andamanese. Andamanese, consisting of the languages spoken on the Andaman Islands, may be a language isolate, but it is believed to be remotely related to the Papuan languages.

Austronesian languages. There are perhaps 500 languages in the Austronesian (Malayo-Polynesian) family, spoken in Malaysia and the Indonesian archipelago; the Philippines; parts of Vietnam, Cambodia, and Taiwan; on the main island groups of the South and Central Pacific; on New Guinea; and on Madagascar. According to one classification, these languages include, in addition to small subgroups, at least two large subgroups: Western Austronesian (or Indonesian) and Eastern Austronesian (often called Oceanic), which includes the Polynesian languages and some of the Melanesian and Micronesian languages. Those Austronesian languages spoken on the Southeast Asian mainland (Malay in Malaysia, Cham and eight other languages mostly in Vietnam, with speakers of some of them also in Cambodia) belong to a Western Indonesian subgroup, which includes Javanese, Sundanese, and Malay, including Bahasa Indonesia, the national language of Indonesia. Closely related to the Western Indonesian subgroup is the subgroup consisting of about 100 languages of the Philippines and a few languages of northern Borneo and northern Celebes (including Tagalog, Cebuano, Hiligaynon, and Ilocano). Classed with the West Indonesian and Philippine languages are a small group of languages of Celebes (*e.g.,* Buginese and Makasarese), a few languages of Borneo, and Malagasy (used on Madagascar).

The languages of Polynesia, including Maori in New Zealand, Tongan, Tahitian, and Hawaiian, form a subgroup that is part of a larger Eastern Oceanic subgroup of more than 100 languages, which includes besides the Polynesian languages such languages as Fijian and a number of languages of the Solomon Islands. At least seven of the languages of Micronesia (including Gilbertese, Chuukese, and Pohnpeian) form another subgroup.

Lingua francas

More than 100 Austronesian languages are spoken throughout New Guinea, and more than 100 Austronesian languages, not counted as Eastern Oceanic, are spoken on smaller islands of Melanesia. Those few with as many as 10,000 speakers are all used as lingua francas in wider areas than those of their native speakers (Dobu in the D'Entrecasteaux Islands, Banoni in southwestern Bougainville, Panayati in the Louisiade Archipelago). Among the Austronesian languages still spoken on Taiwan are Ami, Atayalic, Paiwan, and Bunan. There is some scholarly disagreement concerning the classification of the Austronesian languages (see below *Austronesian languages*).

Other languages. Southeast Asia also has a significant immigrant population that includes speakers of Dravidian languages (Tamil, Telugu, Malayālam, and others) and speakers of Indo-European languages (Punjābī, Bengali, Pashto, and Sinhalese). (See below *Dravidian languages* and *Indo-European languages.*)

NON-AUSTRONESIAN LANGUAGES OF OCEANIA

In effect, the non-Austronesian language areas of New Guinea and Australia together constitute a wedge in the midst of three Austronesian areas: Polynesia to the east, Micronesia to the north, and Indonesia to the west. A few non-Austronesian languages are found on the Indonesian islands nearest to New Guinea (on Halmahera as well as on Timor and Alor).

An exceptionally liberal genetic classification claims that the many non-Austronesian languages in Melanesia and the few in Indonesia all belong to one phylum. Conservative classifications recognize several or even many different language families and avoid the older name for them (Papuan), because it might suggest either that the unrelated families of non-Austronesian languages are branches of one Papuan family or else that non-Austronesian languages are found only on the island of New Guinea. On the other hand, no classification is challenged when it is said that all Australian languages are ultimately related and additionally that they are related neither to Austronesian nor to non-Austronesian languages outside Australia.

In Melanesia, which essentially constitutes the non-Austronesian world beyond Indonesia, there is much contact between Austronesian and non-Austronesian languages. Many of the Melanesian societies are multilingual, especially those in New Guinea; in addition to their native language, speakers often learn a few secondary languages—those of their immediate neighbours or, most frequently, Neo-Melanesian (a pidgin-creole with an English-based lexicon) or both.

In part of Papua New Guinea, Police (or Hiri) Motu, a pidgin based on an Austronesian language, is used as a lingua franca far beyond the territory of the few thousand native speakers of Motu. In Australia the same interest in mastering a multiplicity of languages is widespread, and Aborigines have developed another English-based pidgin-creole, quite different from Neo-Melanesian. Another parallel between Australian languages and the non-Austronesian languages north of Torres Strait is the disinclination of both to recognize or develop any one dialect of a language as a standard.

Papuan languages. About 740 Papuan or non-Austronesian languages extend from the Santa Cruz Islands north and west into the Solomon Islands and the Bismarck Archipelago, across New Guinea to Halmahera, Timor, and Alor. Until the late 1950s all discussions of the languages of New Guinea that treated more than small, closely related groups of languages stressed the fact that the hundreds of languages spoken in a comparatively small area seemed to be completely unrelated to each other except for a few groups of immediate neighbours. Until then, little was known about more than a few of the languages of New Guinea. This situation was changed in the 1960s, with the publication of further survey work in the Highlands region of Papua New Guinea, which stated explicit relationships among a large group of languages. This group was classified the Central New Guinea macrophylum (use of the term *macrophylum* indicates that the languages are less closely related than those of a language family or stock).

Indo-Pacific phylum

There remain a number of families and isolated languages that seem not to be related to other Papuan languages. A liberal classification presented by the American linguist Joseph Greenberg in 1971, however, treats all the Papuan languages as genetically related in an Indo-Pacific phylum, which also includes Andamanese. Most Papuan languages are spoken by only a few hundred to a few thousand speakers (see also *Papuan languages*).

Australian Aboriginal languages. All the Aboriginal languages of Australia are remotely related to each other. A few dozen of the 260 or so Australian languages still spoken account for 90 percent of the total number of speakers. Scores of languages are effectively, if not actually, extinct. The greatest diversity among the languages is found in extreme northern and northwestern Australia (Arnhem Land and Kimberley district); a single remaining family (Pama-Nyungan), with 177 languages, is distributed over the rest of Australia (see below *Australian Aboriginal languages*).

In grammatical typology the non-Austronesian languages north of Torres Strait are heterogeneous, while the Australian languages are syntactically homogeneous and almost identical in patterns of sound combinations. Both Australian languages and non-Austronesian languages have dialects that are linked in a chain such that speakers at either end do not understand the vocabulary of speakers at the other end, although speakers of adjacent dialects can understand each other.

The available data on the two or more languages that were spoken on Tasmania until the later part of the 19th century show a typical Australian sound system, but they have not been linked convincingly to the Australian languages.

LANGUAGES OF AFRICA

Languages that came into Africa from another homeland include, among others, all the European languages associated with 19th-century colonialism. Although the majority of countries in Africa became independent in the 1960s, they continued to use the European languages introduced during the colonial period alongside the numerous languages indigenous to Africa.

Languages from Southwest Asia preceded the languages of European colonization: migrations of peoples to North Africa brought the Ethiopians almost three millennia ago and the Arabic speakers many centuries ago. The Phoenician circumnavigation of Africa in ancient times left traces—Phoenician coins—on the coasts but none in the interior, and long ago migrants from Indonesia reached Madagascar, 250 miles off the African coast. Before and during the colonial period, Arab and Indian traders reached East Africa, where today a few Indo-Aryan languages are spoken among Asians.

The interior of Africa was not known to any non-Africans before the colonial period, but its prehistory can now be partially reconstructed. For example, there is evidence that the homeland of the protolanguage of the numerous Bantu languages was in Cameroon or an adjacent area in West Africa (or in both areas); that a prehistoric migration brought the Bantu speakers to Central and East Africa; and that the movements of these Bantu speakers forced the speakers of San and Khoisan languages to leave their homeland around Lake Victoria and move south to the Kalahari.

Lingua francas

In all the postcolonial nations today, either English or Arabic or French serves both as an international language and as a functioning national language. The question still unresolved for many African nations concerns which of their indigenous languages to develop through writing and to standardize as the official language or languages of education and of the political state. The numerous pidgin-creoles, as Krio, are recent and colonial in inspiration; Sango in the Central African Republic is surely indigenous but not so surely a pidgin-creole. Most of the dozen or so languages used in trade, such as Swahili in East Africa and Hausa in West Africa, tend to have great changes in vocabulary like pidgin-creoles, but they are not classified as pidgin-creoles; instead they are varieties of normal languages that function as lingua francas. Lingua francas of one sort or another are a prerequisite for the markets found throughout rural Africa.

Despite the genetic diversity of the languages of South Africa and the even greater diversity in West Africa, a part of each of these subregions can be shown, on the basis of typology, to be a linguistic area. Thus, most linguists have found that most languages in West Africa distinguish vocabulary items and word elements by tone; in South Africa the clicks characteristic of Khoisan languages also are found among neighbouring Bantu languages such as Xhosa and Zulu. The early use of typology to anticipate genetic classification, however, led to the claim that Africa was full of mixed languages—*e.g.,* Mbugu in Tanzania. But Mbugu, despite having borrowed Bantu prefixes and culture words from Bantu, can be shown to have a single line of origin—to have descended from a single protolanguage (Proto-Cushitic)—on the basis of its grammatical constituents (in particular its pronouns and verb forms) and basic vocabulary items that are cognate with other Cushitic languages.

Afro-Asiatic (Hamito-Semitic) languages. The Hamito-Semitic language family (considered a phylum by some) includes five branches spoken across North Africa from Mauritania to Somalia and beyond into Southwest Asia: Chadic, Semitic, Cushitic, Berber, and the now extinct Egyptian-Coptic. The Chadic branch consists of more than 100 languages spoken in Nigeria, Niger, Cameroon, Ghana, Chad, and the Central African Republic. By far the most widespread is Hausa, estimated to be spoken by as many as 35,000,000 people, for about a third of whom it is a second language.

Five Semitic languages are spoken in Africa, if modern colloquial Arabic is counted as a single language throughout its range across North Africa and the Arabian Peninsula and if Gurage in Ethiopia also is counted as a single language. The Semitic languages in Ethiopia include Amharic, Tigrinya, and Gurage (but the people grouped as Gurage may be speaking several separate languages). Tigré and Tigrinya are spoken in Eritrea.

Cushitic languages are spoken in Eritrea, Ethiopia, Somalia, The Sudan, Tanzania, and Kenya. The languages with the greatest number of speakers are Gallinya, Somali, Sidamo, Hadya, and Afar-Saho. Some scholars consider a group of languages traditionally classified as Cushitic to be a separate branch of Hamito-Semitic, called Omotic. Spoken in Ethiopia, they include Walamo, with far more speakers than the other Omotic languages, Ari, Shako, Zaysse, and others with only a few thousand or a few hundred speakers.

The languages of the Berber branch are spoken from the western desert of Egypt west to the Atlantic and extend to Senegal on the coast and to northern Nigeria in the interior. Guanche, an extinct language that may have been an offshoot of Berber, was formerly spoken on the Canary Islands. Berber languages include Shluh, spoken in Morocco; Tamashek (Tuareg) in Algeria, Libya, Niger, and Mali; and Tamazight in Morocco and Algeria (see below *Hamito-Semitic languages*).

Nilo-Saharan languages. The Nilo-Saharan languages in central interior Africa include the Chari-Nile languages and others that are not closely related to each other or to the Chari-Nile group. (The validity of this grouping has been questioned.) The largest Chari-Nile division, Eastern Sudanic, includes more than 60 languages spoken from Chad to Kenya and Tanzania; it includes a group of languages often classified as a separate family or branch (Nilo-Hamitic), which appears in some classifications as a branch of the Hamito-Semitic family rather than the Nilo-Saharan.

Among the major Eastern Sudanic languages are Teso in Uganda and Kenya, Dinka in The Sudan, Luo in Kenya and Tanzania, and Lango in Uganda. Only three of the 30 or so languages of the Central Sudanic subgroup of Chari-Nile are spoken by groups of some 100,000 people: Sara in Central African Republic and Chad, Lugbara in Uganda and Zaire, and Mangbetu in Zaire.

Among the Nilo-Saharan languages that are not classified as Chari-Nile is the Saharan group. Kanuri, its largest member, is spoken by several million people in Nigeria, Niger, Cameroon, and Chad. In the Maba group, Masalit is spoken in The Sudan. Songhai, often classified as a language isolate, is spoken by about a million people in Niger, Mali, and Burkina Faso. Fur, also sometimes considered to be an isolate, is spoken mostly in The Sudan.

Niger-Congo languages. Languages in the Niger-Congo (or Niger-Kordofanian) family are spoken all across Africa from Mauritania to Kenya and south into South Africa. There are almost 900 Niger-Congo languages, which have been classified into six genetic subgroups. The Bantu (Bantoid) languages of the Benue-Congo subgroup far outnumber those of any other family in Africa, both in terms of number of languages and in terms of total number of speakers. At least 15 Bantu languages are each spoken by more than 3,000,000 people; the following each have more than 5,000,000 speakers: Rwanda, Shona, Kongo, Luba-Lulua (Luba-Kasai), Xhosa, and Zulu.

Bantu languages

Other subgroups in the Niger-Congo family include only a few dozen languages, as those in the Mande subgroup in West Africa, which are spoken from Mauritania to Ghana (including Bambara, Mende, and Vai). The Gur (Voltaic) languages, spoken from Mali and Côte d'Ivoire to Nigeria, include Mossi, with some 4,000,000 speakers, and numerous other languages with significantly fewer speakers. The West Atlantic languages, spoken from Senegal to Nigeria, include Fulani, Wolof, Temne, and several other languages of less numerical import. Of the languages of the Adamawa-Eastern subgroup, spoken from The Sudan to Cameroon, only Sango, through its use as a lingua franca, may be known by more than 1,000,000 people. The Kwa subgroup of Niger-Congo includes Twi (Akan), Yoruba (in Nigeria and Benin), and Igbo (also known as Ibo; in Nigeria). Some scholars link the Kordofanian languages of North and South Kurdufān provinces in The Sudan with the Niger-Congo languages in a Niger-Kordofanian phylum.

Khoisan languages. The Khoisan family consists of about four dozen languages spoken in southern Africa and two click languages (Sandawe and Hadza) spoken in Tanzania that are not closely affiliated with any one group in the Khoisan family. Uncertainties in the number of languages and the number of language groups arise from the profusion of labels for various groups and the lack of detailed linguistic comparisons among large numbers of them. Most of the Khoisan languages have been considered to be on the verge of extinction, if not known to be already extinct, but recent estimates of the numbers of peoples grouped on the basis of their culture (Khoikhoin and San) show many thousands of speakers. The Khoisan language estimated to have the most speakers is Nama. For more information on the Nilo-Saharan (Chari-Nile), Niger-Congo, and Khoisan languages, see below *African languages.*

LANGUAGES OF THE AMERICAS

Languages indigenous to the Americas were brought from Asia by the forebears of modern American Indians (including Eskimos), who left Asia after the dog was domesticated but before other animals were domesticated. Something is known about the culture of these peoples but nothing about their languages, which are known only after contact with European languages.

European languages in the Americas

Today there are six European languages in the Americas that serve as languages of both education and government administration. (Several Indian languages, however, function in this dual role—Guaraní of Paraguay, Greenlandic of Greenland, and Quechua and Aymara of Peru.) These official languages and their number of primary political divisions are Spanish (18); Portuguese (1); Dutch (2)—1 in Latin America and 1 in the Caribbean; English (2 in North America and 11 in the Caribbean); French (1 in North America and 3 in the Caribbean); and Danish (1 in Greenland). Before the colonial period in Latin America and during the first century or two of that period, the following American Indian languages could also be classed as official or semiofficial: Nahuatl (Nahua), the language of the Aztec in Mexico and Central America; Chibcha-Muisca in Colombia; Quechua, the language of the Inca, in the Andean area; Tupí in Brazil; and Guaraní in and around Paraguay. In addition to American Indian languages, two pidgin-creole languages are official in their own political divisions, Sranan (Taki-Taki) in Suriname and Papiamento in Curaçao. Other pidgin-creoles in the Caribbean, such as Haitian Creole, are being increasingly written.

Genetic diversity among languages of continental-sized areas can be expressed in terms of the number of minimum genetic classes taken as the usual basis for discussion by specialists of that area. Research may lead to a downward (or upward) revision, and a new number of minimum genetic classes is used as a basis for further discussion. For North America (north of Mexico) and for the 20th century, the basis for discussion has shifted three times so far: from about 50 families in the classification of the U.S. scholar J.W. Powell to six phyla in the classification of the U.S. anthropological linguist Edward Sapir, which was revised at the 1964 Conference on North American Indian Languages by splitting and reclassification (*e.g.*, of Sapir's Hokan-Siouan) and by merging (*e.g.*, the Muskogean family and a few isolates were added to Algonquian [Algonkian] in the Macro-Algonquian phylum). This third classification is summarized below. Proposals for a minimum number of genetic classes in South America range from more than 100 families to three phyla (in a recent liberal classification).

The Plains Indian sign language (hand talking) is still known, but Chinook Jargon and other pidgin-creoles in North America fell into disuse as soon as American Indians became bilingual in English, French, or Spanish.

North and Central American Indian languages. For North America north of Mexico, the summary of culture areas (before any American Indians were relocated by Europeans) by the U.S. anthropologist Harold E. Driver is a convenient basis on which to superimpose the various ways in which language classifications (genetic and typological) combine with cultures that are ecologically adapted to each of ten areas—the Arctic, Subarctic, Northwest Coast, Plateau, Plains, Prairies, East, California, Great Basin, and Southwest. The three variables (genetic, typological, and cultural) coincide approximately in the Arctic (the one language family, Eskimo-Aleut, does not include typologically diverse languages, but it does spread over a culture area that is not entirely homogeneous). In the Subarctic two language families are represented, Algonkian and Athabascan, which are distinct typologically as well as genetically. Northwest Coast and adjacent Plateau languages are genetically very diversified but surprisingly homogeneous in a diffusional kind of phonological typology. The languages in the treeless Plains and the midwestern Prairies are genetically and typologically diverse; all the language families represented, except Caddoan, are intrusive in the sense that their homelands lie outside the Plains and Prairie areas.

Ten culture areas of North American Indians

Language families in the East give an impression of a little typological similarity combined with considerable genetic diversity. On the opposite coast, California is surprisingly homogeneous in culture and in language typology but heterogeneous in genetic classification of languages. There are few languages and only two language families represented in the Great Basin, which is homogeneous in all respects. The adjacent Southwest is anomalous in all three variables considered here. Where it is culturally homogeneous, as between Pueblo societies, it is genetically and typologically diverse in language: four different language families are represented in Pueblo societies. Non-Pueblo societies of the Southwest are diverse culturally as well as linguistically.

Eskimo-Aleut. The three languages of the Eskimo-Aleut family are still spoken in their prediscovery areas from Greenland to Siberia and also on Komandor Island between the Aleutians and Kamchatka (see below *Languages of the Americas: Eskimo-Aleut languages*).

Athabascan. About 20 languages of the Athabascan family are still spoken in four different culture areas: the Yukon and Mackenzie areas of the Subarctic (the centre of Athabascan diversity, with 17 living languages, including Chipewyan-Slave-Yellowknife and Carrier), the Northwest Coast (where only Hupa, Tolowa, and Chasta Costa may still be spoken), the Southwest (where the Navajo dialect of what may be considered a single Southwestern Apachean language has more speakers than any other Indian language north of Mexico), and the Plains (where two Athabascan languages are more recently intrusive—Sarcee [Sarsi] from the Subarctic and Kiowa Apache from the Southwest). Three language isolates spoken in the Northwest Coast (Eyak, Tlingit, and Haida) are remotely related to Athabascan in the Na-Dené phylum, but Eyak is so much more closely related to the Athabascan family that it might be considered a divergent member of the family.

Algonkian. The Algonkian family includes 13 languages still spoken, which belonged in the culture areas of the eastern Subarctic (*e.g.*, Cree, Ojibwa, Micmac, Malecite), the Prairies (*e.g.*, Fox, Potawatomi), the Plains (*e.g.*, Blackfoot, Cheyenne, Arapaho), and the East (where most Algonkian languages became or are now becoming extinct, with only the removed Shawnee and Delaware surviving in Oklahoma). Remotely related to the Algonkian languages in a Macro-Algonkian phylum are languages spoken further to the south in the East—the Muskogean family (including Choctaw-Chickasaw and Creek-Seminole) and several language isolates that are no longer spoken, as well as two almost extinct languages of the Northwest Coast that are more closely related to Algonkian (Wiyot and Yurok).

Macro-Algonkian phylum

Macro-Siouan. The Macro-Siouan phylum is named for its most extensive component, the Siouan family, the extant languages of which belong in the Plains and Prairies, including Dakota, Crow, Winnebago, and Omaha-Osage. (The Siouan languages of the East, such as Ofo and Biloxi, are no longer spoken.) Less widely distributed than Siouan is the Iroquoian family (six languages, largely of the East, including Cherokee and Mohawk), the Caddoan family (Caddo in the East, Wichita and Pawnee in the Prairies), and two language isolates of the East (Catawba and Yuchi), more closely related to Siouan than to the other families in Macro-Siouan.

Hokan. The Hokan phylum includes several small families and a number of language isolates scattered from the Northwest Coast through California, with extensions into the Great Basin and the Southwest, and as far south as Meso-America. Hokan languages spoken by the greatest numbers of speakers include those in two families in Mexico, the Tlapanecan and Tequistlatecan, and in the Yuman family in Arizona and California.

Penutian. The Penutian phylum is the only group of languages in North America for which relationships with languages in South America have been traced convincingly. The Penutian languages are thus distributed from the Northwest Coast and Plateau areas through California (with a possible extension into the Southwest) and Meso-America into Bolivia, Chile, and Argentina. Many of the Penutian languages north of Mexico are either no longer spoken or are spoken by fewer than 50 people. In Meso-America, however, many native languages have a considerable number of speakers; *e.g.*, Mixe, in the Zoque family, has over 77,000 speakers, and the Mayan family includes some languages with several hundred thousand speakers, as Maya, Quiché, Kekchí, Cakchiquel, and Mam.

Aztec-Tanoan. The Aztec-Tanoan phylum consists of two families: the Tanoan (Kiowa-Tanoan) family with three languages in the Southwest, including those spoken by the Taos and the Santa Clara, and one language in the Plains (Kiowa); and the Uto-Aztecan family, with about a score of languages spoken from the Plateau and California into Meso-America, with relatively late extensions into the Plains. California Uto-Aztecan languages include Cahuilla and Luiseño; Great Basin languages include Paiute and Shoshoni, with the Ute and Comanche dialects in the Plains; Southwestern languages include Hopi and Pima-Papago; Meso-American languages include Nahuatl, the language of the descendants of the Aztecs. The million speakers of the several varieties of Nahuatl far outnumber the total number of speakers of all the other Uto-Aztecan languages.

The Tanoan and Uto-Aztecan families

Oto-Manguean. Languages of one North American phylum are located entirely in Meso-America—the Oto-Manguean phylum, consisting of five small families. The languages with the largest number of speakers are Otomí, Mixtec, and Zapotec.

Unaffiliated languages. In North America one large family (the Salish family in the Northwest Coast and Plateau) and several smaller families and language isolates (as the Wakashan family in the Northwest Coast and Tarascan in Meso-America) remain undetermined in phylum affiliation. Remote relationships that have been proposed for some of these are in conflict with other proposed relationships, with no overwhelming evidence presented for any one of the proposals.

South American Indian languages. Language names for South America are much more numerous than those for North America, but information on actual languages is generally sporadic and often lacking entirely. Even when the list of names is reduced to 350 for languages said to be still spoken, the data to which the names refer consist, for the most part, of brief word lists; or nothing more may be known than the fact that a tribe X is said to speak differently from a tribe Y. Though it is possible to know that certain languages are probably closely related, it is not always possible to say how closely; *i.e.*, whether they might be dialects of, or occasionally just different names for, the same language. At the opposite extreme of genetic relationship, it is clear that there are large groups of remotely related languages, but the paucity of data makes possible conflicting proposals. For at least one group of languages, those of the high cultures of South America—the Inca and the Aymara—and some of their neighbours, the problem of establishing genetic relationship is complicated by the problem of sorting out borrowings among them.

Andean-Equatorial phylum

The Andean-Equatorial phylum includes the greatest number of non-extinct languages (almost 200) and the three South American Indian languages with the greatest number of speakers (Quechua, Guaraní, and Aymara). The living Andean-Equatorial languages constitute some 14 families and several language isolates. The Arawakan family includes the largest number of languages—some 100—and has the widest distribution: across northern South America from French Guiana to Colombia and southward as far as Paraguay; formerly, Arawakan languages also were spoken in Central America and the islands of the Caribbean. Most Arawakan languages are spoken by not more than a few hundred people. More than two dozen languages of the Tupian family are still spoken over a large part of South America, principally south of the Amazon River. Tupian languages include Guaraní (Tupí-Guaraní), which is spoken in a number of dialects by about 4,000,000 people in Paraguay, Brazil, Argentina, and Bolivia. Quechua, of the Quechumaran group, is spoken by some 8,000,000 people in Peru, Ecuador, Colombia, Bolivia, Argentina, and Chile. Some Quechua dialects are so divergent that they might be regarded as separate languages. The other Quechumaran language group, Aymaran, is spoken by more than 1,000,000 people in Peru and Bolivia. Most other languages in the Andean-Equatorial phylum are spoken by only a few thousand persons.

The Ge-Pano-Carib phylum includes almost as many languages still spoken as the languages of the Andean-Equatorial phylum, but the former are all spoken by relatively small tribes, so that the total number of speakers of these languages is only a small fraction of the number of speakers of Andean-Equatorial languages. In terms of numbers of languages, the largest family in the Ge-Pano-Carib phylum is the Cariban (Carib) family, with some 60 languages still spoken in Venezuela, French Guiana, Guyana, Suriname, Brazil, and Colombia. Cariban languages were also formerly spoken in the Caribbean islands. Most Cariban languages have fewer than 1,000 speakers. The other large family in the phylum, the Macro-Ge family, includes more than 25 languages in Brazil.

The languages of the Macro-Chibchan phylum, of which some 39 may still be spoken, are distributed from Guatemala and Honduras southward into, and possibly beyond, Peru. The largest component of the phylum is the Chibchan family, of which 16 languages are still spoken from Nicaragua to northwestern Colombia—these include Cuna, spoken on the San Blas Archipelago of Panama as well as on the mainland of Panama and Colombia; Guaymí in Panama; and Páez in Colombia.

For further information on the Indian languages of the Americas, see below *Languages of the Americas: North American Indian languages; Meso-American Indian languages; South American Indian languages.*

For information on numbers of speakers by country, see the *Britannica World Data: Language* section in the BRITANNICA BOOK OF THE YEAR. (C.F.V./F.M.V./Ed.)

INDO-EUROPEAN LANGUAGES

Indo-European is the name of a family of languages that by 1000 BC were spoken over most of Europe and in much of Southwest and South Asia; since the second half of the 15th century the Indo-European languages have spread to most other inhabited parts of the world. The term Indo-Hittite is used by scholars who believe that Hittite and the other Anatolian languages (see below) are not just one branch of Indo-European but rather a branch coordinate with all the rest put together; thus, Indo-Hittite has been used for a family consisting of Indo-European proper plus Anatolian. As long as this view is neither definitively proved nor disproved, it is convenient to keep the traditional use of the term Indo-European.

Overview of the language family

LANGUAGES OF THE FAMILY

The well-attested languages of the Indo-European family fall fairly neatly into the 10 main branches listed below; these are arranged according to the age of their oldest sizable texts.

Anatolian. Now extinct, Anatolian was spoken during the 1st and 2nd millennia BC in what is presently Asian Turkey and northern Syria. By far the best-known of its members is Hittite, the official language of the Hittite empire, which flourished in the 2nd millennium. Very few Hittite texts were known before 1906, and their interpretation as Indo-European was not generally accepted until after 1915; the integration of Hittite data into Indo-European comparative grammar has, therefore, been one of the principal developments of Indo-European studies in the 20th century. The oldest Hittite texts date from the 17th century BC, the latest from approximately 1200 BC. For more information, see below *Anatolian languages.*

Indo-Iranian. Indo-Iranian comprises two main subbranches, Indo-Aryan (Indic) and Iranian. Indo-Aryan languages have been spoken in what is now northern and central India and Pakistan since before 1000 BC. Aside from a very poorly known dialect spoken in or near northern Iraq during the 2nd millennium BC, the oldest record of an Indo-Aryan language is the Vedic Sanskrit of the Rigveda (Ṛgveda), the oldest of the sacred scriptures of India, dating roughly from 1000 BC. Examples of modern Indo-Aryan languages are Hindī, Bengali, Sinhalese (spoken in Sri Lanka), and the many dialects of Romany, the language of the Gypsies (Rom).

Iranian languages were spoken in the 1st millennium BC in present-day Iran and Afghanistan and also in the steppes to the north, from modern Hungary to East (Chinese) Turkistan. The only well-known ancient varieties of Iranian languages are Avestan, the sacred language of the Zoroastrians (Parsis), and Old Persian, the official language of Darius I (ruled 522–486 BC) and Xerxes I (486–465 BC) and their successors. Among the modern Iranian languages are Persian (Fārsī), Pashto (Afghan), Kurdish, and Ossetic. For more information, see below *Indo-Iranian languages.*

Greek. Greek, despite its numerous dialects, has been a single language throughout its history. It has been spoken in Greece since at least 1600 BC, and, in all probability, since the end of the 3rd millennium. The earliest texts are the Linear B tablets, some of which may date from as far back as 1400 BC (the date is disputed), and some of which certainly date to 1200 BC. This material, very sparse and difficult to interpret, was not identified as Greek until 1952. The Homeric epics—the *Iliad* and the *Odyssey*—probably dating from the 8th century BC, are the oldest texts of any bulk. For more information, see below *Greek language.*

Linear B texts

Italic. The principal language of the Italic group is Latin, originally the speech of the city of Rome and the ancestor of the modern Romance languages: Italian, Romanian, Spanish, Portuguese, French, and so on. The earliest Latin inscriptions apparently date from the 6th century BC, with literature beginning in the 3rd century. Scholars are not in agreement as to how many other ancient languages of Italy and Sicily belong in the same branch as Latin. For more information on Latin, the languages derived from it, and the other languages that belong to or are sometimes included in the Italic branch of Indo-European, see below *Italic languages* and *Romance languages.*

Germanic. In the middle of the 1st millennium BC, Germanic tribes lived in southern Scandinavia and northern Germany. Their expansions and migrations from the 2nd century BC onward are largely recorded in history. The oldest Germanic language of which much is known is the Gothic of the 4th century AD. Other languages include English, German, Dutch, Danish, Swedish, Norwegian,

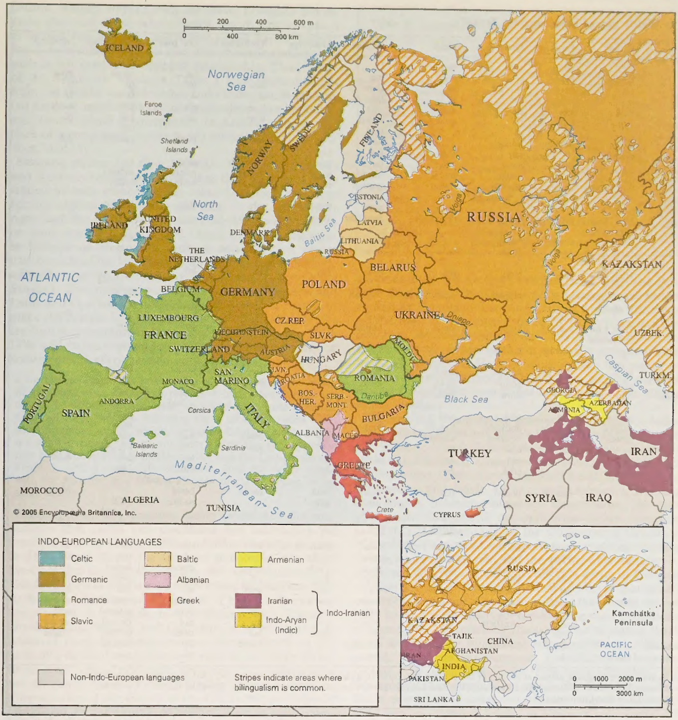


Figure 1: Approximate locations of Indo-European languages in contemporary Eurasia.

and Icelandic. For more information, see below *Germanic languages* and *English language*.

Armenian. Armenian, like Greek, is a single language. Speakers of Armenian are recorded as being in what now constitutes eastern Turkey and Armenia as early as the 6th century BC, but the oldest Armenian texts date from the 5th century AD. For more information, see below *Armenian language*.

Tocharian. The Tocharian languages, now extinct, were spoken in the Tarim Basin (in present-day northwestern China) during the 1st millennium AD. Two distinct languages are known, labeled A (East Tocharian, or Turfanian) and B (West Tocharian, or Kuchean). One group of travel permits for caravans can be dated to the early 7th century, and it appears that other texts date from the same or from neighbouring centuries. These languages became known to scholárs only in the first decade of the 20th century; they have been less important for Indo-European studies than has Hittite, partly because their testimony about the Indo-European parent language is obscured by 2,000 more years of change and partly because Tocharian testimony fits fairly well with that of the previously known non-Anatolian languages. For more information, see below *Tocharian languages*.

Celtic. Celtic languages were spoken in the last centuries before the Christian era over a wide area of Europe, from Spain and Britain to the Balkans, with one group (the Galatians) even in Asia Minor. Very little of the Celtic of that time and the ensuing centuries has survived, and this branch is known almost entirely from the Insular Celtic languages—Irish, Welsh, and others—spoken in and near the British Isles, as recorded from the 8th century AD onward. For further information, see below *Celtic languages*.

Celtic in mainland Europe

Balto-Slavic. The grouping of Baltic and Slavic into a single branch is somewhat controversial, but the exclusively shared features outweigh the divergences. At the beginning of the Christian Era, Baltic and Slavic tribes occupied a large area of eastern Europe, east of the Germanic tribes and north of the Iranians, including much of present-day Poland and what was formerly the western Soviet Union—namely, Belarus, Ukraine, and westernmost Russia. The Slavic area was in all likelihood relatively small, perhaps centred in what is now southern Poland. But in the 5th century AD the Slavs began expanding in all directions. By the early 21st century the Slavic languages were spoken throughout much of eastern Europe and northern Asia. The Baltic-speaking area, however, contracted, and by the early 21st century Baltic languages were confined to Lithuania and Latvia.

The earliest Slavic texts, written in a dialect called Old Church Slavonic, date from the 9th century AD; the oldest substantial material in Baltic dates to the end of the 14th

Table 1: Widely Shared Indo-European Terms*

	Hittite	Sanskrit	Greek	Latin	English	Armenian	Tocharian B	Old Irish	Lithuanian	Albanian
'I'	uk	ahám	egṓ	ego	I	es			àš	
'me'	ammuk	mā́m	emé	mē	me	is	ñäś	-m	manè	mua
'thou'	zik	tuvám	sú, tú	tū	thou	du	twe	tú	tù	ti
'thee'	tuk	tvā́m	sé	tē	thee	k'ez	ci	-t	tavè	ty
'who?'	kuiš	kás	tís	quis	who?	ov	k_use	cía	kàs	kush
'what?'	kuit	kím	tí	quid	what?	z-i	k_use	cid	kàs	çë
'that'		tát	tó		that	da	te		taĩ	
'water'	wātar	udakám	húdōr		water		war	uisce	vanduõ	ujë
'fire'	paḫḫur		pũr	Umbr. pir	fire	hur	puwar			
'father'		pitár-	patḗr	pater	father	hayr	pācer	athair		
'mother'		mātár	mḗtēr	māter	mother	mayr	mācer	máthair	mótina	
'brother'		bhrā́tar-	phrā́tēr 'kinsman'	frāter	brother	ełbayr	procer	bráthair	brólis	
'sister'		svásār		soror	sister	k'oyr	ṣer	siur	seser-	
'daughter'	Hier. Luw. tuwatra-	duhitár	thugátēr-	Osc. futír	daughter	dustr	tkācer		dukter-	
'son'		sūnús	huiós		son		soy		sūnùs	
'sheep'	Luw. ḫawi-	ávis	ó(w)is	ovis	ewe	hovi-w 'shepherd'	awi 'ewes'	oí	avìs	
'cow'	Hier. Luw. wawi-	gā́v-	boũs	bōs	cow	kov	ke_u	bó	Latv. gùovs	
'horse'	Hier. Luw. aśuwa-	áśvas	híppos	equus	OE eoh		yakwe	ech	ašvà 'mare'	
'pig'		sūkarás	hũs	sūs	sow		suwo			thi
'dog'	Hier. Luw. śuwana-	śvā́n-	kúōn	canis	hound	šun	kwen-	con-	šun-	
'wheel'		cakrám	kúklos		wheel		kokale 'wagon'			
'heart'	kart-	hṛ́dayam	kardíā	cord-	heart	sirt		cride	širdìs	
'knee'	gēnu	jā́nu	gónu	genū	knee	cunr	keni	glún		gju
'tree, wood'	tāru	dā́ru	dóru	truncus 'tree trunk'	tree		or	daur 'oak'	OCS drěvo	dru
'foot'	pat(a)-	pā́d-	pód-	ped-	foot	otn	paiyye		pãdas	
'long'	dalukiš	dīrghás	dolikhós	longus	long				ìlgas	
'new'	newaš	návas	ne(w)os	novus	new	nor	ñuwe	nue	naũjas	
'goes'	pa-izzi	éti	eĩsi	it			yan		OL eĩti	
'is'	ēšzi	ásti	estí	est	is	ē		is	ėsti	është
'eats'	ēzz(a)zzi	átti	édei	ēst	eats	utē		estir 'may eat'	ėda	
'carries'		bhárati	phérei	fert	bears	berē	parän	berid		bie 'brings'
'knows'		véda	(w)oĩde		wot	gitē		ro-fitir	OCS věstŭ	
'one'		ékas	oĩ(w)os 'alone'	ūnus	one			oín	víenas	një
'two'	dā-	duvā́	dúo	duo	two	erku	wi	dó	dù	dy
'three'	teri-	tráyas	treĩs	trēs	three	erek'	trey	trí	trỹs	tre
'four'		catvā́ras	téttares	quattuor	four	č'ork'	śtwer	cethair	keturì	katër
'five'		páñca	pénte	quīnque	five	hing	piś	cóic	penkì	pesë
'six'		ṣáṭ	héx	sex	six	vec'	ṣkas	sé	šešì	gjashtë
'seven'	šiptam-	saptá	heptá	septem	seven	ewt'n	ṣukt	secht	septynì	shtatë
'eight'		aṣṭā́	oktṓ	octō	eight	ut'	okt	ocht	aštuonì	tetë
'nine'		náva	enné(w)a	novem	nine	inn	ñu	noí	devynì	nëndë
'ten'		dáśa	déka	decem	ten	tasn	śak	deich	dẽšimt	dhjetë
'one hundred'		śatám	he-katón	centum	hundred		kante	cét	šim̃tas	
'not'	natta	ná		ne-	not			ní-	ne-	

*Words lacking in the language named at the top of the column but found in a closely related language are included, with these abbreviations: Luw. = Luwian; Hier. Luw. = Hieroglyphic Luwian; OE = Old English; Latv. = Latvian; OCS = Old Church Slavonic; Umbr. = Umbrian; Osc. = Oscan; OL = Old Lithuanian.

century, and the oldest connected texts to the 16th century. For more information, see below *Baltic languages* and *Slavic languages.*

Albanian. Albanian, the language of the present-day republic of Albania, is known from the 15th century AD. It presumably continues one of the very poorly attested ancient Indo-European languages of the Balkan Peninsula, but which one is not clear. For more information, see below *Albanian language.*

In addition to the principal branches just listed, there are several poorly documented extinct languages of which enough is known to be sure that they were Indo-European and that they did not belong in any of the groups enumerated above (*e.g.,* Phrygian, Macedonian). Of a few, too little is known to be sure whether they were Indo-European or not (*e.g.,* Ligurian).

ESTABLISHMENT OF THE FAMILY

Shared characteristics. The chief reason for grouping the Indo-European languages together is that they share a number of items of basic vocabulary, including grammatical affixes, whose shapes in the different languages can be related to one another by statable phonetic rules. Especially important are the shared patterns of alternation of sounds. Thus the agreement of Sanskrit *ás-ti,* Latin *es-t,* and Gothic *is-t,* all meaning 'is,' is greatly strengthened by the identical reduction of the root to *s-* in the plural in all three languages: Sanskrit *s-ánti,* Latin *s-unt,* Gothic *s-ind* 'they are.' Agreements in pure structure, totally divorced from phonetic substance, are, at best, of dubious value in proving membership in the Indo-European family.

Shared vocabularies

Table 1 gives examples of typical vocabulary items widely shared within the Indo-European family that have been decisive in establishing the family. A blank indicates that the language in question does not use the item in accordance with the given meaning or that its word for that meaning is unknown.

Similarities in grammatical endings are shown in Table 2 by samples of noun declension and verb inflection in some of the more archaic languages that have retained the inflectional endings of Indo-European in relatively unchanged form. Note that Old Lithuanian *-į* and *-ų* were nasalized vowels, representing a continuation from the earlier forms **-in* and **-un.* (The asterisk marks a form that is not actually found in any document or living dialect but is reconstructed as having once existed in the prehistory of the language.)

The statable phonetic rules referred to earlier are not always obvious without careful observation. Note that the English dental consonants *t, d,* and *th* do not correspond in a straightforward manner to the Greek dental sounds *t, d,* and *th;* that is, English *t* does not occur where Greek *t* appears, nor English *d* where Greek has *d.* But the relationships between the sounds are not random either—English *t* does not correspond to Greek *t* in one word, to *d* in a second, and to *th* in a third, according to no discernible pattern. Rather, where Greek has initial *t*, English has *th*, as in *that* and *three;* where Greek has *d*, English has *t*, as in *tree, two,* and *ten;* and where Greek has *th*, English has *d*, as in *daughter.* Note also that phonetic similarity as such is not needed to establish relationship. Thus, many of the Armenian words in Table 1 look quite different from the related words in other Indo-European languages, but here too regular rules of correspondence can be found; *e.g.,* Greek initial *p* corresponds to Armenian *h* or zero (lack of a consonant) in the words meaning 'fire,' 'father,' 'foot,' and 'five.'

Sanskrit studies and their impact. The ancient Greeks and Romans readily perceived that their languages were related to each other, and, as other European languages became objects of scholarly attention in the late Middle Ages and the Renaissance, many of these were seen to be more similar to Latin and Greek than, for example, to Hebrew or Hungarian. But an accurate idea of the true bounds of the Indo-European family became possible only when, in the 16th century, Europeans began to learn Sanskrit. The massive similarities between Sanskrit and Latin and Greek were noted early, but the first person to make the correct inference and state it conspicuously was the British Orientalist and jurist Sir William Jones, who in 1786 said in his presidential address to the Bengal Asiatic Society that Sanskrit bore to both Greek and Latin

> a stronger affinity, both in the roots of verbs, and in the forms of grammar, than could possibly have been produced by accident; so strong, indeed, that no philologer could examine them all three, without believing them to have sprung from some common source, which, perhaps, no longer exists. There is a similar reason, though not quite so forcible, for supposing that both the *Gothick* [*i.e.,* Germanic] and the *Celtick,* though blended with a very different idiom, had the same origin with the *Sanscrit;* and the old *Persian* might be added to the same family . . . .

Nineteenth-century linguists firmly established the connections that Jones had elucidated and broadened the family to include Slavic, Baltic, and other language groups. In 1816 Franz Bopp, the German philologist, presented his *Über das Conjugationssystem der Sanskritsprache in Vergleichung mit jenem der griechischen, lateinischen, persischen und germanischen Sprache* ("On the System of Conjugation in Sanskrit, in Comparison with Those of Greek, Latin, Persian, and Germanic"), in which the relation of these five languages was demonstrated on the basis of a detailed comparison of verb morphology (structure). Two years later there appeared the *Undersøgelse om det gamle Nordiske eller Islandske Sprogs Oprindelse* (*Investigation of the Origin of the Old Norse or Icelandic Language*), by the Danish philologist Rasmus Rask, completed in 1814. This work demonstrated methodically the relation of Germanic to Latin, Greek, Slavic, and Baltic. (Rask included Celtic a few years later.) In 1822 the second edition of the first volume of Jacob Grimm's *Deutsche Grammatik* ("Germanic Grammar") was published; in this grammar were discussed the peculiar Indo-European vowel alternations called *Ablaut* by Grimm (*e.g.,* English "sing, sang, sung"; or Greek *peíth-ō* 'I persuade,' *pé-poith-a* 'I am persuaded,' *é-pith-on* 'I persuaded'). In addition, Grimm tried to find the principle behind the correspondences of Germanic stop and spirant consonants (the first made with complete stoppage of the breath, and the second made with constriction of the breath but not complete stoppage) to the consonants of other Indo-European languages. The sound changes implied by these correspondences have become known as Grimm's law. An example of Grimm's law is the correspondence of the stop consonant *p* in Latin *pater* with the spirant consonant *f* in *father.*

The work of Bopp, Rask, and Grimm

Bopp demonstrated in 1839 that the Celtic languages were Indo-European, as had been asserted by Jones. In 1850 the German philologist August Schleicher did the same for Albanian, and in 1877 another German philologist, Heinrich Hübschmann, showed that Armenian was an independent branch of Indo-European, rather than a member of the Iranian subbranch. Since then, the Indo-European family has been enlarged by the discovery of Tocharian and of Hittite and the other Anatolian languages. The Indo-European character of Tocharian was announced by the German scholars Emil Sieg and Wilhelm Siegling in 1908. The Norwegian Assyriologist Jørgen Alexander Knudtzon recognized Hittite as Indo-European on the basis of two letters found in Egypt (translated in *Die zwei Arzawa-briefe* [1902; "The Two Arzawa Letters"]), but his views were not generally accepted until 1915, when Bedřich Hrozný published the first report of his own decipherment of the much more copious material that had meanwhile been found in the ruins of the Hittite capital itself.

Table 2: Examples of Noun and Verb Inflection

	Hittite pānt- ('gone')	Sanskrit yant- ('going')	Greek iónt- ('going')	Latin eunt- ('going')	Old Lithuanian seser- ('sister')
Singular nominative	pānz(a)	yán	iṓn	iēn-s	sesuõ
Singular accusative	pānt-an	yánt-am	iónt-a	eunt-em	sẽser-į
Singular genitive	pānt-aš	yat-ás	iónt-os	eunt-is	seser-ès
Singular dative	pānt-i	yat-é		eunt-ī	sẽser-i
Singular locative	pānt-i	yat-í	iónt-i	eunt-e	seser-yjè
Plural nominative	pānt-eš	yánt-as	iónt-es	eunt-ēs	sẽser-es
Plural accusative	pānt-uš	yat-ás	iónt-as	eunt-ēs	sẽser-is
Plural genitive	pānt-an	yat-ā́m	iónt-ōn	eunt-(i)um	seser-ų̃
I go	pai-mi	é-mi	eĩ-mi	e-ō	ei-mì
You (sg.) go	pai-ši	é-ṣi	eĩ	ī-s	ei-sì
He, she goes	pai-zzi	é-ti	eĩ-si	i-t	eĩ-ti
We go	pai-wani	i-más	í-men	ī-mus	ei-mè
You (pl.) go	pai-tteni	i-thá	í-te	ī-tis	ei-tè
They go	pa-anzi	y-ánti	í-āsi	e-unt	

Bopp's comparative Indo-European grammar

The first full comparative grammar of the major Indo-European languages was Bopp's *Vergleichende Grammatik des Sanskrit, Zend, Griechischen, Lateinischen, Litthauischen, Altslawischen, Gotischen und Deutschen* (1833–52; "Comparative Grammar of Sanskrit, Zend, Greek, Latin, Lithuanian, Old Slavic, Gothic, and German"). But this and August Schleicher's shorter *Compendium der vergleichenden Grammatik der indogermanischen Sprachen* (1861–62; "Compendium of the Comparative Grammar of the Indo-European Languages") were rendered obsolete by the major breakthrough of the 1870s, when scholars—prompted largely by the discoveries of a group of German scholars known as Neogrammarians—realized that sound correspondences are not merely rules of thumb that do not have to be strictly observed, but that apparent exceptions to sound laws can often be accounted for by stating them more accurately or by reconstructing additional different sounds in the parent language. The difference between Gothic *d* in *fadar* 'father' and *þ* in *broþar* 'brother,' for example, both corresponding to *t* in Sanskrit, Greek, and Latin, proved to be correlated with the original position of the accent, a discovery known as Verner's law (named for the Danish linguist Karl Verner). Thus, *d* appears when the preceding syllable was originally unaccented (*fadar* : Greek *patér-*, Sanskrit *pitár-*), and *þ* occurs when the preceding syllable was originally accented (*broþar* : Greek *phráter-* 'member of a clan,' Sanskrit *bhrátar-*).

The knowledge and opinions that had accumulated by the end of the 19th century are largely incorporated in the German linguist Karl Brugmann's *Grundriss der vergleichenden Grammatik der indogermanischen Sprachen* (2nd ed., 1897–1916; "Outline of Comparative Indo-European Grammar"), which remains the latest full-scale treatment.

THE PARENT LANGUAGE: PROTO-INDO-EUROPEAN

By comparing the recorded Indo-European languages, especially the most ancient ones, much of the parent language from which they are descended can be reconstructed. This reconstructed parent language is sometimes called simply Indo-European, but in this article the term Proto-Indo-European is preferred.

Phonology. *Consonants.* Proto-Indo-European probably had 15 stop consonants. In the following grid these sounds are arranged according to the place in the mouth where the stoppage was made and the activity of the vocal cords during and immediately after the stoppage:

	labial	dental	palatal	velar	labiovelar
Voiceless	p	t	ḱ	k	k^w
Voiced	b	d	ǵ	g	g^w
Voiced aspirated (?)	bh	dh	ǵh	gh	g^wh

A "labial" sound is made with the lips, and a "dental" sound with the tip of the tongue against the back of the teeth. The "palatal" and "velar" sounds were probably made by contact between the back of the tongue and the soft palate—more toward the front of the mouth in the case of the palatals and more toward the back in the case of the velars (compare Arabic _kalb_ 'dog' versus _qalb_ 'heart'). The "labiovelar" sounds were made by contact between the back of the tongue and the soft palate with concomitant rounding of the lips. "Voiceless" designates sounds made without vibration of the vocal cords; "voiced" sounds are pronounced with vibration of the vocal cords. The exact pronunciation of the "voiced aspirates" is somewhat uncertain; they were probably similar to the sounds transcribed _bh, dh,_ and _gh_ in Hindī.

Correspondences pointing to the voiced labial stop _b_ are rare, leading some scholars to deny that _b_ existed at all in the parent language. A minority view holds that the traditionally reconstructed voiced stops were actually glottalized sounds produced with accompanying closure of the vocal cords. The status of the velar stops _k, g,_ and _gh_ has likewise been questioned. The earlier view that Proto-Indo-European had a series of voiceless aspirated stops _ph, th, ḱh, kh,_ and k^wh has largely been abandoned. There was one sibilant consonant, _s,_ with a voiced alternant, _z,_ that occurred automatically next to voiced stops. The existence of a second apical spirant, _þ_ (presumed pronunciation like that of _th_ in English _thin_), is extremely uncertain.

Laryngeal theory

There is general agreement that Proto-Indo-European had one or more additional consonants, for which the label "laryngeal" is used. These consonants, however, have mostly disappeared or have become identical with other sounds in the recorded Indo-European languages, so that their former existence has had to be deduced mainly from their effects on neighbouring sounds. Hence, the laryngeal sounds were not suspected until 1878, and even then they were rejected by most scholars until after 1927, when the Polish linguist Jerzy Kuryłowicz showed that Hittite often has _ḫ_ (perhaps a velar spirant like the _ch_ in German _ach_) in places where a laryngeal had been posited on the evidence of the other Indo-European languages. There is still considerable disagreement about how many laryngeals there were, what they sounded like, what traces they left, and how best to symbolize them. Most scholars now believe there were three, which can be written H_1, H_2, and H_3. Of these, H_1 may have been _h_ or a glottal stop; H_2 was perhaps a pharyngeal spirant like Arabic _ḥ_ in _ḥams_ 'five'; H_3, whatever its other features, was probably voiced. The principal traces they left outside Anatolian are in the quality and length of neighbouring vowels, H_2 changing a neighbouring _e_ to _a,_ and probably H_3 changing it to _o,_ while all laryngeals lengthened a preceding vowel in the same syllable. In Anatolian, H_2 and H_3 remained as _ḫ,_ at least in some positions.

When laryngeals between consonants disappeared, a vowel sometimes remained, as in Greek _stásis,_ Sanskrit _sthitis,_ Old English _stede_ 'a standing (place)' from Proto-Indo-European _*stH_2tis._ Before the advent of the laryngeal theory, a separate Proto-Indo-European vowel _ə_ (called _schwa indogermanicum_) was reconstructed to account for these correspondences.

Finally, there were the nasal sounds _n_ and _m,_ the liquids _l_ and _r,_ and the semivowels _y_ and _w._ When _y_ and _w_ occurred between consonants, they were replaced by the vowels _i_ and _u._ The nasals and liquids functioning as nuclei of syllables in this position (like the final sounds of English _bottom, button, bottle, butter_) are traditionally written _n̥, m̥, l̥, r̥._ Some scholars dispense with these diacritical marks and with the distinction between syllabic _i_ and _u_ and nonsyllabic _y_ and _w,_ but this obscures certain distinctions, such as that between _-wn̥-_ in _*ḱwn̥su_ 'among dogs,' Sanskrit _śvasu,_ and _-un-_ in _*tund-_ 'shove,' Sanskrit _tundate._

Vowels. The vowel system of Proto-Indo-European consisted of the following sounds:

	front			back
high	i, ī			u, ū
mid		e, ē	o, ō	
low		a, ā		

In forming front vowels, the highest point of the tongue is in the front of the mouth; for back vowels, that point is in the back. High vowels are those in which the tongue is highest—closest to the roof of the mouth; mid vowels are made with the tongue between the extremes of high and low.

Ablaut

The four mid vowels participated in a pattern of alternation called "ablaut." In the course of inflection and word formation roots and suffixes could appear in the "_e_-grade" (also called "normal grade"; compare Latin _ped-is_ 'of a foot' [genitive singular]), "_o_-grade" (_e.g.,_ Greek _pód-es_ 'feet'), "zero-grade" (_e.g.,_ Avestan _fra-bd-a-_ 'fore-foot,' with _-bd-_ from _*-pd-_), "lengthened _e_-grade" (_e.g.,_ Latin _pēs_ 'foot' [nominative singular] from _*pēd-s_), and/or "lengthened _o_-grade" (_e.g.,_ English _foot,_ Old English _fōt_).

There is some evidence for a similar pattern of alternation involving _a, ā,_ and zero. Most instances of apparent _a_ and _ā,_ however, arose by "coloration" of _e_ under the influence of a preceding or following H_2 (_e.g.,_ Greek _ag-_ 'lead' comes from _*H_2eǵ-,_ _stā-_ 'stand' comes from _*stH_2-_). Some cases of _o, ō,_ and _ē_ are likewise of laryngeal origin (_e.g.,_ Greek _op-_ 'see' comes from _*H_3ek^w-,_ _dō-_ 'give' comes from _*deH_3-,_ _thē-_ 'put' comes from _*dheH_1-_). Among the high vowels, _i_ and _u_ did not participate in ablaut alternations but rather functioned primarily as the syllabic realizations of the consonants _y_ and _w,_ as in _*leyk^w-_ 'leave,' zero-grade _*lik^w-,_ parallel to _*derḱ-_ 'see,' zero-grade _*dr̥ḱ-._ Long _ī_ and _ū_ in the recorded languages derive in large part from sequences of _i_ or _u_ plus laryngeal, as in Latin _vīvus_ 'alive' from _*g^{w}iH_3wós._

The accent just before the breakup of the parent language was apparently mainly one of pitch rather than stress. Each full word had one accented syllable, presumably pronounced on a higher pitch than the others.

Morphology and syntax. _Verbal inflection._ The Proto-Indo-European verb had three aspects: imperfective, perfective, and stative. Aspect refers to the nature of an action as described by the speaker—_e.g.,_ an event occurring once, an event recurring repeatedly, a continuing process, or a state. The difference between English simple and "progressive" verb forms is largely one of aspect—_e.g.,_ "John wrote a letter yesterday" (implying that he finished it) versus "John was writing a letter yesterday" (describing an ongoing process, with no implication as to whether it was finished or not).

The imperfective aspect, traditionally called "present," was used for repeated actions and for ongoing processes or states—_e.g.,_ _*stí-stH_2-(e)-_ 'stand up more than once, be in the process of standing up,' _*mn̥-yé-_ 'ponder, think,' _*H_1es-_ 'be.' The perfective aspect, traditionally called "aorist," expressed a single, completed occurrence of an action or process—_e.g.,_ _*steH_2-_ 'stand up, come to a stop,' _*men-_ 'think of, bring to mind.' The stative aspect, traditionally called "perfect," described states of the subject—_e.g.,_ _*ste-stóH_2-_ 'be in a standing position,' _*me-món-_ 'have in mind.'

Verb roots were by themselves either perfective (like _*steH_2-_ 'stand' and _*men-_ 'think') or imperfective (like _*H_1es-_ 'be'). This basic aspect, however, could be reversed by morphological devices such as ablaut, suffixation, and reduplication. The stative aspect was normally marked by reduplication and the _o_-grade of the root in the indicative singular; it had personal endings that were partly distinct from those of the other two aspects.

From one aspect of a given verb the shape and even the existence of the other two aspects could not be predicted; for example, _*H_1es-_ 'be' had only the imperfective aspect. Ways of forming imperfectives were especially numerous and often involved, in addition to their imperfective aspectual meaning, some other notion, such as performing the action habitually or repeatedly (iterative), or causing someone else to perform it (causative). One root could thus have several imperfective stems; so to the root _*H_1er-_ 'move' there were at least a causative form, _*H_1r̥-new-_ 'set in motion,' and an iterative form, _*H_1r̥-sḱe-_ 'go repeatedly.'

The Proto-Indo-European verb was also inflected for mood, by which the speaker could indicate whether he was making statements or inquiries about matters of fact; making predictions, surmises, or wishes about the future or about unreal but imagined situations; or giving commands. Compare English "If John is home now (he is eating lunch)" with the verb *is* in the indicative mood, discussing a matter of fact, with "If John were home now (he would be eating lunch)" with the verb *were* in the subjunctive mood, describing an unreal situation. There were two Proto-Indo-European suffixes expressing mood: -e- alternating with -o- for the subjunctive, corresponding roughly in meaning to the English auxiliaries 'shall' and 'will,' and -yeH_1- alternating with -iH_1- for the optative, corresponding roughly to English 'should' and 'would.' Verbs without one of these two suffixes were marked for mood and tense by their personal endings alone.

Mood in Proto-Indo-European

These personal endings basically expressed the person and number of the verb's subject, as in Latin *amō* 'I love,' *amās* 'you (singular) love,' *amat* 'he or she loves,' *amāmus* 'we love,' and so on. In the imperfective and perfective aspects there were two sets of endings, distinguishing two voices: active, in which typically the subject was not affected by the action, and mediopassive, in which typically the subject was affected, directly or indirectly. Thus Sanskrit active *yájati* and mediopassive *yájate* both mean 'he sacrifices,' but the former is said of a priest who performs a sacrifice for the benefit of another, while the latter is said of a layman who hires a priest to perform a sacrifice for him. In the stative aspect there was originally no distinction of voice.

To mark mood and tense, imperfective verbs that did not have a mood suffix distinguished three subtypes of active and mediopassive endings: imperative, primary, and secondary. Verbs with imperative endings belonged to the imperative mood (used for commands)—*e.g.*, *H_1s-dhí 'be (singular),' *H_1és-tu 'let him be.' Verbs with primary endings were marked as non-past (present or future) in tense and indicative in mood—*e.g.*, *H_1és-ti 'he is.' (Indicative mood signifies objective statements and questions.) Verbs with secondary endings were unmarked for tense and mood but were normally used as past indicatives (*e.g.*, *H_1és-t 'he was,' *g^whén-t 'he slew') and to fill out gaps in the imperative paradigm (*e.g.*, *H_1és-te or *H_1s-té 'you [plural] were,' but also 'be [plural]'; *g^whén-te or *g^whn̥-té 'you [plural] slew,' but also 'slay [plural]'). To mark such forms unambiguously as past indicatives, an augment, usually consisting of the vowel *e*, could be prefixed—*e.g.*, *é-g^when-t 'he slew,' *é-H_1es-t 'he was.'

Verbs in the perfective aspect without a mood suffix did not occur with primary endings and thus lacked a true present tense. Verbs in the stative aspect substituted a distinctive set of endings for those of the primary set but apparently used the imperative and secondary endings in the usual way to form a stative imperative and a stative past indicative.

Nominal inflection. The inflectional categories of the noun were case, number, and gender. Eight cases can be reconstructed: nominative, for the subject of a verb; accusative, for the direct object; genitive, for the relations expressed by English *of;* dative, corresponding to the English preposition *to,* as in "give a prize to the winner"; locative, corresponding to *at, in;* ablative, *from;* instrumental, *with;* and vocative, used for the person being addressed. For examples of some of these see Table 2. Besides singular and plural number, there was a dual number for referring to two items. Each noun belonged to one of three genders: masculine, to which belonged most nouns designating male creatures; feminine, to which belonged most names of female creatures; and neuter, to which belonged only a few words for individual adult living creatures. The gender of nouns not designating living creatures was only partly predictable from their meaning.

Gender

Adjectives were nounlike words that varied in gender according to the gender of another noun with which they were in agreement, or, if used by themselves, according to the sex of the entity to which they referred; thus, Latin *bonus sermō* 'good speech' (masculine), *bona aetās* 'good age' (feminine), *bonum cor* 'good heart' (neuter), or *bonus* 'a good man,' *bona* 'a good woman,' *bonum* 'a good thing.' The neuter of an adjective was often identical with the masculine except for having different endings in the nominative and accusative cases. Feminine gender was either completely identical with the masculine or derived from it by means of a suffix, the two commonest being *-eH_2- and *-iH_2- (*-yeH_2-).

Demonstrative, interrogative, relative, and indefinite pronouns were inflected like adjectives, with some special endings. Personal pronouns were inflected very differently. They lacked the category of gender, and they marked number and case (in part) not by endings but by different stems, as is still seen in English singular nominative "I," but oblique "my," "me"; plural nominative "we," but plural oblique "our," "us." (The oblique is any case other than nominative or vocative.)

Syntax. Some notable features of Proto-Indo-European syntax were the non-ergative case system, in which the subject of an intransitive verb received the same case marking as the subject (rather than the object) of a transitive verb; concord (agreement) in case, number, and gender between adjective and noun; and the use of singular verbs with neuter plural subjects, as in Greek *pánta rheî* 'all things flow,' with the same (singular) verb as *ho pótamos rheî* 'the river (masculine) flows,' contrasting with *hoi pótamoi rhéousi* 'the rivers flow' (indicating that neuter plurals were originally collectives and grammatically singular). Proto-Indo-European word order was flexible, but basic declarative sentences typically had the structure subject–object–verb (SOV).

Lexicon and culture. Much less is known about the parent language's vocabulary than about its phonology and grammar. Sounds and grammatical categories do not easily disappear or undergo radical change in so many daughter languages that their former existence can no longer be detected. It is relatively easy, however, for an individual word to disappear or shift meaning in so many daughter languages that its existence or meaning in the parent language cannot be confidently inferred. Hence, from the linguistic evidence alone, scholars can never say that Proto-Indo-European lacked a word for any particular concept; they can only state the probability that certain items did exist and from these items make inferences about the culture and location in time and space of the speakers of Proto-Indo-European.

Inferences from vocabulary

Thus is it supposed that the Proto-Indo-European community knew and talked about dogs (*ḱwón-), horses (*H_1éḱwo-), sheep (*H_3éwi-), and almost certainly cows (*g^wṓw-) and pigs (*súH-). Probably all these animals were domesticated. At least one cereal grain was known (*yéwo-), and at least one metal (*H_2éyos). There were vehicles (*wóǵho-) with wheels (*k^wék^wlo-), pulled by teams joined by yokes (*yugó-). Honey was known, and it probably formed the basis of an alcoholic drink (*mélit-, *médhu) related to the English *mead.* Numerals up through 100 (*ḱm̥tóm) were in use. All this suggests a people with a well-developed Neolithic (characterized by simple agriculture and polished stone tools) or even Chalcolithic (copper- or bronze-using) technology.

The divergence of Indo-European languages. Linguists have not found a reliable and precise way to determine from linguistic evidence alone the date at which any set of related languages must have begun diverging. The best that can be done is to estimate the degree of difference between the languages in question, taking into account all that is known about them, and then compare this estimate with the estimated degrees of difference within families of languages—such as the Romance family—whose actual time of divergence is approximately known. Using this sort of "dead reckoning," it can be said that the earliest attested Indo-European languages—Anatolian, Indo-Iranian, and Greek—are different enough that the parent language must have been split into several distinct languages before 3000 BC, but similar enough that the first split into separate languages is not likely to have been earlier than about 4500 BC.

For further progress the linguistic findings must be correlated with archaeological evidence. Linguistic, historical, and geographic considerations suggest that the speakers of Proto-Indo-European were a relatively small and homogeneous Eurasian population group that underwent significant expansion and fragmentation in the period around 4000 BC. Some scholars believe that the Indo-Europeans were the bearers of the Kurgan (Barrow) culture of the Black Sea and the Caucasus and west of the Urals.

The Kurgan culture, however, was only one of a number of related steppe cultures extending across the entire Black Sea–Caspian Sea region, an area that was transformed about 4000 BC by the advent of horse-drawn wheeled vehicles and related innovations. It is probably best, therefore, to follow J.T. Mallory (*In Search of the Indo-Europeans* [1989]) in locating the speakers of Proto-Indo-European among the populations of this region, but not to attempt a more precise identification until further evidence is available.

Possible relationship to Uralic

Remote relationship of Indo-European to the Uralic languages is not improbable. Geographically, the earliest reconstructible locations of the two families are contiguous; lexically, there are strong resemblances in a number of basic words or word parts, including personal, demonstrative, interrogative, and relative pronouns, personal endings of verbs, the accusative case ending *-m,* and such words as those for 'water' and 'name'; typologically, the families are fairly similar—*e.g.,* both have many suffixes, but few or no prefixes or infixes (elements inserted within words). The resemblances, however, are too few to permit the reconstruction of a common "Indo-Uralic" parent language; the two families, if they are related at all, must have separated thousands of years before the breakup of Proto-Indo-European.

If Indo-European is related to other language families—*e.g.,* to Afro-Asiatic (which includes the Semitic languages) or to Kartvelian (which includes Georgian)—it must have diverged from them much earlier than it diverged from Uralic, because the number of cogent resemblances is much smaller. There is no significant evidence at present for a "Nostratic" superfamily embracing these and other groups.

CHARACTERISTIC DEVELOPMENTS OF INDO-EUROPEAN LANGUAGES

As Proto-Indo-European was splitting into the dialects that were to become the first generation of daughter languages, different innovations spread over different territories.

Changes in phonology. Indo-Iranian, Balto-Slavic, Armenian, and Albanian agree in changing the palatal stops *ḱ, *ǵ, and *ǵh into spirants (*s, ś, th,* etc.) or affricates—*e.g.,* Sanskrit *aśri-* 'sharp edge,' Old Church Slavonic *ostrŭ* 'sharp,' Armenian *asełn* 'needle,' Albanian *athëtë* 'bitter' beside Greek *ắkros* 'tip,' Latin *acidus* 'biting,' all from a basic element $*H_2ek$- 'sharp, pointed.' (Spirants, also called fricatives, are sounds produced with audible friction as a result of the airstream passing through a narrow, but unstopped, passage in the mouth—*e.g.,* English *s, f, v.* Affricates are sounds that begin as stops, with complete stoppage of the airstream, but are released as spirants, or fricatives—*e.g.,* the *ch* in *church,* the *j* in *jam.*) The languages that change the palatal stops to spirants or affricates are known as "satem" languages, from the Avestan word *satəm* 'hundred' (Proto-Indo-European *ḱm̥tóm), which illustrates the change. The languages that preserve the palatal stops as *k*-like sounds are known as "centum" languages, from *centum* (/kentum/), the corresponding word in Latin. The satem languages are not geographically separated from one another by any recorded languages that preserve the palatals as stops; it is therefore inferred that the change to affricates (whence later spirants) occurred just once and spread over a cohesive dialect area of Proto-Indo-European.

Of the languages that share this change, however, Balto-Slavic shares with Germanic (including English) an *m* in certain case endings where other Indo-European languages, including Indo-Iranian, Armenian, and Albanian, have *bh* or a sound regularly developed from *bh.* Examples of the *m* ending include English *the-m* and Old Church Slavonic *tě-mŭ* 'to those ones'; the *bh* and related sounds (*ph, v, b*) are illustrated in the following: Sanskrit *té-bhyas* 'to those ones,' Armenian *noro-vkʿ* 'with new ones,' Albanian *male-ve* 'to mountains,' Greek *ṓkhes-phin* 'with chariots,' Latin *omni-bus* 'for all.' Because Balto-Slavic and Germanic are neighbours, it is inferred that *m* replaced *bh* in these case endings just once in the parent language and that the area over which this innovation spread only partly overlapped the area that adopted affricated pronunciation of the palatals.

This pattern is general for changes dating from the time the parent language was breaking up into distinct languages. Each of the resulting languages shares some innovations with some of its neighbours, but only rarely do different innovations shared by two or more branches of Indo-European cover exactly the same territory.

Developments of the separated Indo-European languages

Once the dialects had become differentiated enough to be distinct languages—certainly by 2500 BC in most cases—each largely went its own way, and agreements in developments since then are due either to borrowing across language boundaries (as in the notable convergences between Modern Greek, Albanian, Romanian, and the southernmost Slavic languages) or to parallel but independent workings out of the same base material.

In phonology, the most striking changes have been loss or reduction in many languages of final or unaccented syllables, and loss in several languages of certain consonants between vowels, often followed by contraction of the resulting vowel sequence. Thus words in modern Indo-European languages are often much shorter than their Proto-Indo-European ancestors—*e.g.,* English 'four,' Armenian *čʿorkʿ,* colloquial Persian *čar* 'four' from $*k^w et$-*wóres;* French *vit* (pronounced *vi*) 'lives' from $*g^w íH_3 weti$; Russian *dvestí* 'two hundred' from $*duwóyH_1$ $ḱm̥tóyH_1$.

Changes in morphology. As a result of the fact that much of the marking of Proto-Indo-European inflectional categories was done in final syllables, loss and reduction of these syllables have often had serious grammatical consequences. In the noun, loss of endings has generally led to loss or great reduction of the case and gender systems, while ways have generally been found to salvage the distinction between singular and plural. In Modern Persian, for example, where all final syllables have been lost, the old case and gender distinctions have disappeared also, but plural number is still regularly marked, either with *-an* (originally the genitive plural ending of some nouns) or with *-ha* (of obscure origin).

In the verb, where more endings originally had two syllables, loss of final syllables has had less serious consequences for morphology. Even here, however, some languages, including English, have totally or almost totally given up the marking of subject by personal endings. Compare English "I, we, you, they love" and "he, she loves" with the Spanish conjugation for 'love'—*amo, amas, ama, amamos, amáis, aman*—or the Russian version—*ljubljú, ljúbish, ljúbit, ljúbim, ljúbite, ljúbjat.*

Changes in noun inflection have generally involved simplification. Almost everywhere the dual number has been lost; in many languages the noun genders have been reduced from three to two (as in French, Swedish, Lithuanian, and Hindī) or lost entirely (as in English, Armenian, and Bengali). Only Slavic has complicated the gender system by imposing on the inherited distinctions contrasts of animate versus inanimate or of personal versus nonpersonal.

Everywhere except in the oldest Indo-Iranian languages the original eight Indo-European cases have suffered reduction. Proto-Germanic had only six cases, the functions of ablative (place from which) and locative (place in which) being taken over by constructions of preposition plus the dative case. In Modern English these are reduced to two cases in nouns, a general case that does duty for the vocative, nominative, dative, and accusative ("Henry, did Bill give John the letter?") and a possessive case continuing the old genitive ("Bill's letter"). In languages such as French and Welsh, nouns are no longer inflected for case at all. In some languages, to be sure, nouns have begun fusing with words placed directly after the nouns to create new case systems, coexisting with relics of the old. Thus, Old Lithuanian had in addition to seven inherited cases an illative (place into), made by adding *-n(a)* to the accusative (*peklosna* 'into hell'), an allative (place to, toward), made by adding *-p(i)* to the genitive (*Jesausp* 'to Jesus'), and an adessive (place at which), made by adding *-p(i)* to the locative (*Joniep* 'in John').

Changes in the verb

Changes in the verb have been more complex. Besides loss or merger of old categories, many new forms have been created and many old forms have acquired new values. In Ancient Greek the focus of the stative aspect (perfect) has largely shifted from the present state ("he is dead") to the previous event that led to this state ("he has died"). As a result, the perfect came to mean the same as the perfective past (aorist), and it has therefore disappeared from Modern Greek. New forms created in Ancient Greek include future and future perfect tenses, based on the desiderative present forms (such as "he wants to walk") of the parent language.

In Germanic the principal new creation was the weak past tense (ending in a *t* or *d*), such as English *loved, thought,* German *liebte, dachte,* made by combining the verb stem with a past tense of the Germanic verb for 'do.' (The strong past tense formed by vowel alternations, like "sing, sang," "run, ran" comes from the Proto-Indo-European stative aspect.)

In some languages participles have come to function as finite verbs. Thus in Hindī *ādmī laṛkī-ko dekhtā* 'the man sees the girl,' *dekhtā* 'sees' is etymologically a participle 'seeing,' agreeing in number and gender with the subject *ādmī* 'man.' In the past tense, *ādmī-ne laṛkī dekhī* 'the man saw the girl,' the verb *dekhī* is etymologically a past passive participle 'seen,' agreeing in gender and number with the object *laṛkī* 'girl,' and the subject is marked with an instrumental ending.

Vocabulary changes. Changes in vocabulary have been even greater than those in sounds and grammar. Words in modern Indo-European languages have several sources. They may be recognizable loanwords, such as English *skunk, chain,* and *inch* (from Algonquian, French, and Latin, respectively); they may have been formed within the history or prehistory of the language itself, such as English *radar* and *rightness;* they may be of obscure origin, such as English *drink,* which is common Germanic but has no cognates outside Germanic, or *boy,* which is peculiar to English and Frisian; or they may be inherited words that have changed meaning, such as English *merry* from Proto-Indo-European **mr̥ǵhú-* 'short.' Only a small fraction of the vocabulary can be traced back to words that can confidently be asserted to have existed in the parent language with approximately their present meaning. The same is true, albeit in a lesser degree, even for the oldest recorded Indo-European languages. None has more than a few hundred words and roots that are clearly inherited from the parent language without essential change of meaning. Table 1 gives examples of words that have been widely retained with little change. Typically they include pronouns; nouns, verbs, and adjectives of relatively simple and ubiquitous meaning; numerals; and simple adverbs and prepositions.

Non-Indo-European influence on the family. The influence of non-Indo-European languages on the sounds and grammar of Proto-Indo-European is not demonstrable, partly because there is no direct evidence about the languages that were in contact with Indo-European before roughly 3000 BC. It can be surmised, however, that some words are loans—*e.g., *pélek̑u-* 'ax,' a word for an object likely to be imported or learned of from neighbours with superior technology and which is not analyzable into a known Indo-European root plus a known Indo-European suffix.

Reliance on linguistic conjecture

When Indo-European languages have been carried within historic times into areas occupied by speakers of other languages, they have generally taken over a number of loanwords, as with English and Spanish in the Americas or Dutch in South Africa. Aside from the special case of pidgin and creole languages, however, there has been comparatively little effect on sounds and grammar. These have been significantly affected within historic times only when an Indo-European language has been spoken in prolonged close contact with non-Indo-European speakers, as with Ossetic (an Iranian language) in the Caucasus, or when its speakers have been very strongly influenced culturally by speakers of a non-Indo-European language, as with Persian, in which Arabic plays much the same role as Latin does in English.

In prehistoric times most branches of Indo-European were carried into territories presumably or certainly occupied by speakers of non-Indo-European languages, and it is reasonable to suppose that these languages had some effect on the speech of the newcomers. For the lexicon, this is indeed demonstrable in Hittite and Greek, at least. It is much less clear, however, that these non-Indo-European languages affected significantly the sounds and grammar of the Indo-European languages that replaced them. Perhaps the best case is India, where certain grammatical features shared by Indo-European and Dravidian languages appear to have spread from Dravidian to Indo-European rather than vice versa. For most other branches of Indo-European languages any attempt to claim prehistoric influence of non-Indo-European languages on sounds and grammar is rendered almost impossible because of ignorance of the non-Indo-European languages with which they might have been in contact. (W.C./J.H.Ja.)

Anatolian languages

The term Anatolian languages in its most comprehensive use includes both the Indo-European and non-Indo-European languages spoken in Anatolia (Asia Minor) before the Greco-Roman period. The Anatolian languages are known only from texts of the 2nd and 1st millennia BC; the earliest evidence is that of the so-called Cappadocian tablets (19th–18th centuries BC). The term Asianic is sometimes used as an alternative designation for the Anatolian languages, but, since the discovery in 1915 that Hittite, the main Anatolian language, is an Indo-European language, there has been a tendency to use Asianic in a more restricted sense for the non-Indo-European languages that existed in Anatolia before the entry of the Indo-Europeans. These are called substratum languages.

Hattic (or Hattian), also misleadingly called Proto-Hittite, is the best-known substratum language. It is completely unrelated to Hittite and its sister languages as well as to Hurrian, a language also spoken in Anatolia.

The Anatolian group of Indo-European languages consists of Hittite, Palaic, Luwian, Hieroglyphic Luwian, Lydian, and Lycian. Hittite, Palaic, and Luwian are known from 2nd-millennium cuneiform texts found in the excavations in Boğazköy-Hattusa since 1905; Hieroglyphic Luwian is found on scattered inscriptions and seals from Anatolia (mainly the southern area) and northern Syria dating mainly from later times (*i.e.,* between *c.* 1200 and 700 BC, although there are earlier examples from the empire period, *c.* 1400–*c.* 1190 BC). Lydian and Lycian are known from texts in alphabetic script from *c.* 600 to 200 BC. It seems fairly reasonable to add the Carian language of southwest Anatolia to this list as well as other less well documented languages like Sidetic. More to the east, in the Caucasus region centring around Lake Van, Hurrian of the 3rd and 2nd millennia BC was replaced in the 1st millennium BC by the related Urartian language. Both of these languages are definitely non-Indo-European.

Historical background of ancient Anatolia. It is customarily assumed that the Indo-Europeans entered Anatolia around or shortly after 2000 BC, although there are no specific archaeological data that might enable scholars to specify the period of entry or the route the invaders followed. On the basis of the agricultural terminology used in Hittite, it has been suggested that the entry into Anatolia was not a warlike invasion of predominantly male groups. If such had been the case, the influence of substratum languages would have been likely, but, on the contrary, the word stems used are definitely Indo-European. The differences in the terminology used in other Indo-European subgroups indicate that the "Anatolians" seceded from the parent group at an early date, before the common agricultural nomenclature came into being. On the other hand, Hittite shares the Indo-European notion of the hereafter, pictured as a pastureland with grazing cattle "for which the dead king sets out."

Indo-European entry into Anatolia

There is a tendency among linguists to postulate an eastern route of entry into Anatolia by way of the Caucasus, because certain grammatical features—*e.g.,* the loss of the feminine gender—might be explained as having been caused by prolonged contacts with Caucasian languages.

It is likely that the Indo-European forebears of the later speakers of Hittite, Palaic, Luwian, and Lydian entered Anatolia together, following a common route, because the Anatolian languages share a considerable number of losses as well as innovations that presuppose a long common past.

In the central parts of Anatolia, within the bend of the Halys River (modern Turkish, Kizil Irmak), and in the northern regions, Hittite and Palaic were profoundly influenced by Hattic as a substratum language. The Hattian culture also changed the political and religious concepts of the newcomers, and a clear cultural dependency of the Indo-Europeans on the older Hattian population is evident. Some scholars have stressed the likelihood that farther to the south the Luwians might have been conversant with a different substratum. In view of the absence of textual evidence, and because knowledge of the Luwian vocabulary is rather restricted, it is perhaps not surprising that this possible substratum element escapes definition. (For the history of Anatolia in the 2nd and 1st millennia BC, see TURKEY AND ANCIENT ANATOLIA.)

The most important invaders of Anatolia in the "Dark Age" (after 1190 BC) were the Phrygians. Their language is definitely Indo-European, but it bears no relationship to the Anatolian subgroup. Rather, it seems akin to Thracian, Illyrian, or possibly Greek. Greek, in the second half of the 1st millennium BC, and, later, Latin, from the 2nd century onward, entered central Anatolia as languages of a ruling caste. Much earlier—beginning in Mycenaean times—the west coast had attracted Greek settlers. In the first half of the 1st millennium, the southern and northern shores also attracted Greek-speaking peoples. To the east in the Caucasus region, other Indo-Europeans, the Armenian-speaking invaders, penetrated into the former Urartian territory well before the beginning of the Persian period, probably in the 7th and 6th centuries BC. During Persian times, a Persian ruling caste entered eastern and also northeastern Anatolia and was still clearly recognizable in the Hellenistic and Roman periods (*e.g.*, in Bithynia, Pontus, Cappadocia, and Commagene). Late data on names and scattered remarks made by Fathers of the Church indicate that until late Roman and perhaps even Byzantine times, some Anatolian dialects remained in use in certain isolated parts of the interior.

Early research on Anatolian languages

Classification of the languages. Research on the Anatolian languages began in 1821 with the Lycian language and passed an initially fruitful phase in the 1880s with work on Hieroglyphic Hittite (nowadays referred to as Hieroglyphic Luwian). In 1902 the Norwegian Assyriologist Jørgen Alexander Knudtzon's study on the Arzawa letters was published; these were two letters exchanged between a king of Arzawa and Pharaoh Amenhotep III that had been found in the Amarna archive. They were written in the Hittite language in cuneiform writing. In 1915 research reached a climax with the interpretation of Cuneiform Hittite by the Czech Orientalist Bedřich Hrozný. In all four of these highlights, the discovery that the texts in question were Indo-European was either clearly expressed or more discreetly implied. This conclusion was based on both the nominal (noun) declension and the verbal conjugation: the languages had a nominative ending in *-s*, the accusative in *-n*, verbal endings like *-ti* and *-nti* for the 3rd person singular and plural of the present tense, and an imperative form like *estu* "let it be." These features were deemed to be sufficient proof of their Indo-European origin. Study of the Anatolian subgroup of Indo-European thus began with Lycian, the last Anatolian offshoot in the temporal sequence, then passed the intermediary stage of Hieroglyphic Luwian, and reached the 2nd-millennium Hittite language in 20th-century research. For the relationship between members of the Anatolian subgroup, see Figure 2.

The non-Indo-European Hurrian and Urartian languages are related to one another, but modern research indicates that Urartian should not be considered as a direct continuation of Hurrian.

HISTORY AND DEVELOPMENT

Languages using cuneiform writing and Anatolian hieroglyphs. *Hattic.* The Hattic language appears as *hattili* in Hittite cuneiform texts. Called Proto-Hittite by some, it was the language of the linguistic substratum inside the Halys River bend and in more northerly regions. Apparently the Indo-European newcomers of Hittite stock were named with the same designation as their predecessors. All the Hattic material preserved by Hittite scribes belongs to the religious sphere of life: rituals (*e.g.*, connected with the erection of a new building), incantations, antiphons, litanies, and myths. Among the Hattic interpolations in Hittite texts, there are some to which a Hittite translation has been added. It is impossible to ascertain the length of time that the Hattians had been present in Anatolia before the Indo-Europeans entered the country, but it seems certain that during the Hittite New Empire (*c.* 1400–*c.* 1190 BC) Hattic was a dead language.

Records of Hattic

Hattic studies began in 1922 with the work of the German Assyriologist Emil Forrer. In 1935, Hans G. Güterbock, a German-born Orientalist, published a large group of texts containing Hattic material and in so doing completed the publication of the Hattic texts stemming from the Winckler excavations (1905–12). Important studies on the subject have continued to appear since then.

Hittite. The Hittite language is known from the approximately 25,000 tablets or fragments of tablets preserved in the archives of Boğazköy-Hattusa, excavated by German archaeologists beginning in 1905. In Hittite cuneiform texts, the language is referred to as *nesili* (*nasili*) "language of Nesa," or *nesumnili* "language of the Neshite." Earlier Hittite linguistic material may be found in the indigenous proper names and a few loanwords from the local dialect that are recorded in the Cappadocian tablets (the commercial correspondence in Assyrian of Assyrian colonists living in Anatolia, especially in the emporium at Kültepe, near modern Kayseri, between *c.* 1900 and 1720 BC). The data from Kültepe are sometimes referred to as "Kaneshite" (from Kanesh, the old name of Kültepe); this is obviously the modern equivalent of the word *kanisumnili* "language of the Kaneshite" found in a Hittite text. It is possible, or even likely, that Kanesh and Nesa do, in fact, refer to the same entity.

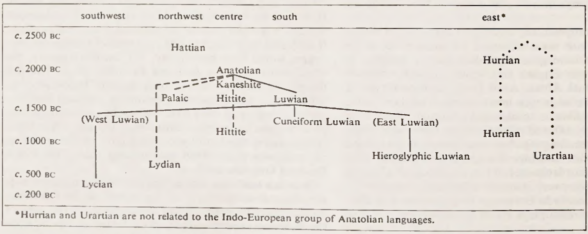


Figure 2: Relationship between members of the Anatolian subgroup.

Hittite tablets from places outside of the Hittite capital are rare; only stray examples have been found—*e.g.*, in Tarsus, Alalakh, Ugarit, and Amarna. These findings attest to the growth of a great Hittite empire, especially between *c.* 1400 and *c.* 1190 BC. Old Hittite, the written embodiment of the earliest Indo-European language that has been discovered so far, is known from some tablets preserved in an "old ductus" type of handwriting that was typical of copies from the Old Kingdom period (*c.* 1700–1500 BC). The intermediary "Dark Age" between *c.* 1500 and *c.* 1400 BC is sometimes referred to as the period of the so-called Middle Hittite language. Most of the Old and Middle Hittite texts, however, are preserved in copies from the later empire period.

Archives of Boğazköy-Hattusa

The archives of Boğazköy-Hattusa have been found in various places in the citadel, in the Great Temple complex, and in the "House on the Slope." Although the majority of the texts are concerned with religious subjects (oracle texts, hymns, prayers, myths, rituals, and festival texts), these archives also contain material of historical, political, administrative, literary, and legal character. The cuneiform adopted by the Hittite scribes is a variant of a writing system of Mesopotamian origin that closely resembles the ductus and shapes prevalent in tablets of the 17th century BC (layer VII) from Alalakh (modern Atsana in southeastern Turkey). It is possible that the cuneiform script might have been introduced as a result of the Hittites inducing Syrian scribes to transfer their activities to the Hittite capital during the early part of the Old Kingdom, shortly after 1650 BC. It has also been posited, with good reason, that the newly acquired script was first used to write Akkadian and was only later employed for Hittite as well. In addition to the genres enumerated above, the "scholarly literature" deserves to be mentioned. This consists of the material considered by the scribes to be essential for their training; it includes word lists, omens, and ritual prescriptions, all reflecting an encyclopaedic approach aimed at complete coverage of the subjects concerned. The Sumerian texts found in these archives belong to this class of literature. For treaties and correspondence with foreign powers, Akkadian was used as the diplomatic language of that period. Therefore, both Sumerian and Akkadian formed part of the curriculum of the qualified scribes, these languages belonging to the "eight languages" found in the Hittite archives.

In actual fact, the first decipherer of Hittite was the Norwegian scholar J.A. Knudtzon, who pointed out in 1902 that the language of the so-called Arzawa letters (*i.e.*, Hittite)—found in the Amarna archive—had an apparent affinity with Indo-European. Because the cuneiform script had already been deciphered, Knudtzon, and Bedřich Hrozný after him, were able to "read" their texts. Thus their discovery consisted more in the interpretation than in the actual decipherment of the written material. The first series of German excavations, lasting from 1905 to 1912, produced about 10,000 tablets. It was work on this corpus that familiarized Hrozný with the contents of these tablets and led him to his epoch-making discovery that Hittite was indeed Indo-European (1915).

Palaic. Palaic, which appears as *Palaumnili* "language of the Palaite" in Hittite cuneiform texts, was the language of the region of Pala (probably Blaëne in the Greek period), in northwest Anatolia. During the Old Hittite kingdom, Pala, Luwiya, and Hattusa formed the three major provinces of the Anatolian part of the Hittite territory. From the intermediary "Dark Age" onward, Kaska nomads made their influence felt in northern Anatolia, and this resulted in a decline of importance for this region.

Linguistic affiliations of Palaic

The Indo-European character of Palaic was first advocated by Emil Forrer (1922). Part of the text material is preserved on tablets in "old ductus." The knowledge of the limited vocabulary leaves much to be desired, but parallels—especially in the inflection of the noun, the forms of the demonstrative, relative, and enclitic pronouns, and the verbal endings—vouch for a close relationship to Hittite and Luwian.

Luwian. Luwian (or Luvian), the language of Anatolia's southern coast, is known from texts stemming from three major periods: (1) the Hittite New Empire (*c.* 1400–*c.* 1190 BC); (2) the period of the Neo-Hittite states (*c.* 1190–*c.* 700 BC); (3) the period of the Lycian monumental inscriptions (*c.* 400–200 BC). In addition to the various time periods, there is also a variation in writing system—Mesopotamian cuneiform, Anatolian hieroglyphs, and an alphabet derived from a Greek source—and dialectal differentiation. There are indications that as early as the 15th and 14th centuries BC, there was a West Luwian dialect (the precursor of alphabetic Lycian) and an East Luwian dialect (the forerunner of the later Hieroglyphic Luwian of the Neo-Hittite states). Both of these differed from the Luwian found in the archives of Boğazköy-Hattusa, which was possibly a central dialect.

As in the case of Palaic, the pioneering work on Luwian written in cuneiform was done by Emil Forrer (1922). Following this work, new text materials were published in 1953, closely followed by both grammatical and vocabulary studies as well as a standard dictionary of Cuneiform Luwian (1959).

The Anatolian hieroglyphic system has a long history, with its logographic beginnings dating back to early Hittite stamp seals of the 18th and 17th centuries BC; the youngest texts seem to date from the last quarter of the 8th century BC. The geographical range of the inscriptions is great, stretching from Sipylus and Karabel in the extreme west to Alaca Hüyük and Boğazköy-Hattusa in the north, Malatya, Samsat, and Tell Ahmar (Til Barsib) in the east, and Hama and ar-Rastān in the south. During the "Dark Age" of the 16th and 15th centuries BC, the early writing grew into a fully developed writing system with logograms (word-signs), syllabic values, and auxiliary signs. During the New Empire, the script was already in use for a multitude of purposes (rock inscriptions, seals, and wooden tablets for everyday use in the temple and the army). Whether an example of the empire period such as the Aleppo inscription already reflects the Luwian language is a moot question but seems likely. It is certain that the later inscriptions of the Neo-Hittite states were in Luwian.

Work on Hieroglyphic Luwian

The first attempts to decipher Hieroglyphic Luwian, made by the British archaeologist Archibald H. Sayce, were fortunate in some fundamental details, but it was not until the 1930s that systematic and mutually stimulating research by scholars of several countries led to the establishment of a number of syllabic values for the characters as well as to a correct analysis of the sentence structure of the inscriptions. In his publication of the (bilingual) Hittite royal seals (in 1940, 1942), Hans G. Güterbock bridged the gap between the inscriptions of the empire period and the late Neo-Hittite states; the seals found in the French excavations at Ugarit (in northern Syria) served a similar purpose. The most important recent finding was the discovery in 1947 by Helmuth T. Bossert, a German archaeologist, of the Karatepe bilingual inscriptions, written in Phoenician and Hieroglyphic Luwian.

On many points the Luwian vocabulary is still an enigma. The unity between the various Luwian dialects and the close relationship of Luwian to the other members of the Anatolian subgroup, however, is secured by several linguistic parallels, especially in the singular inflection of the noun, the forms of certain pronouns, the verbal endings, and a number of lexical (vocabulary) correspondences.

Hurrian. In earlier stages of research, the terms Mitanni language and Subarian were used as designations for Hurrian. In Hittite cuneiform texts, *hurlili* "language of the Hurrian" is used. In the last centuries of the 3rd millennium BC, Hurrians were already present in the Mardin region, which, from a geographical point of view, belongs to the North Mesopotamian plain. In Mesopotamian texts (from the time of the Akkad dynasty) some Hurrian personal names and glosses have been found. The customary assumption is that this non-Semitic and also non-Indo-European ethnic group had come from the Armenian mountains. During the beginning of the 2nd millennium BC, the Hurrians apparently spread over larger parts of southeast Anatolia and northern Mesopotamia. Still later, during the intermediary "Dark Age," they are supposed to have infiltrated into Cilicia and the adjacent Taurus and Antitaurus regions (Kizzuwatna in 2nd millennium texts). Before the middle of the 2nd millennium BC, an

Indo-Aryan ruling caste wielded some type of authority over parts of Hurrian territory. Some names and words in ancient Near Eastern texts bear witness to their presence. Among these words are a group of technical terms related to the training of horses that found its way into Hittite treatises on that subject; they are most important from a historical point of view. After Sumerian, Akkadian, Hattic, Palaic, and Luwian, Hurrian and these Indo-Aryan glosses constitute the sixth and seventh additional languages of the Hittite archives.

Sources of Hurrian texts

Hurrian texts have been found in Urkish (Mardin region, *c.* 2300 BC), Mari (on the middle Euphrates, 18th century BC), Amarna (Egypt, *c.* 1400 BC), Boğazköy-Hattusa (Empire period), and Ugarit (on the coastline of northern Syria, 14th century). Amarna yielded the most important Hurrian document, a political letter sent to Pharaoh Amenhotep III. From Mari came a small number of religious texts; from Boğazköy-Hattusa, literary and religious texts; and from Ugarit, vocabularies belonging to the more "scholarly literature" described above and Hurrian religious texts in Ugaritic alphabetic script. Hurrian personal names, found in texts from many sites (Boğazköy-Hattusa, Alalakh, Ugarit, and especially Nuzu), constitute a second linguistic source of major importance.

The research on Hurrian started in the 1890s with simultaneous contributions by several scholars. Subsequently, Bedřich Hrozný (1920) and Emil Forrer (1919, 1922) discovered the presence of Hurrian material in the Boğazköy-Hattusa archives.

Urartian. The terms Chaldean and Vannic have also been used as designations for Urartian during earlier stages of research. Urartian is not a late dialect of Hurrian but a separate language, although both stem from a common parent. During the 9th through 6th centuries BC, Urartian was used in northeastern Anatolia as the official language of the state of Urartu, which centred around the district of Lake Van but also extended over the Transcaucasian regions of modern Russia and into northwestern Iran and at times even into parts of North Syria. The Urartian texts are written in a variant of the Neo-Assyrian script and consist mostly of monumental inscriptions (annals, votive inscriptions related to building and irrigation activities), some small inscriptions on helmets and shields dedicated in the temple, and a few economic cuneiform tablets. Two bilingual inscriptions in Urartian and Assyrian that apparently correspond very closely provided the key to the understanding of the language; the stylistic resemblances to Assyrian texts of the same period guided the further interpretation.

Archibald H. Sayce was the first scholar to devote his attention to Urartian in the 1880s and 1890s and continued his activities until 1932. More important were the philological contributions of the German historian Carl F. Lehmann-Haupt between 1892 and 1935. The first reliable description of Urartian grammar was published by the German Orientalist Johannes Friedrich (1933).

Next to the Urartian texts in cuneiform writing, there also existed an indigenous hieroglyphic script that is still undeciphered and is too meagrely represented to warrant a serious attempt.

Languages using a derivative of the Phoenician or Greek alphabet. *Phrygian.* The Phrygian inscriptions and graffiti may be separated into two groups, the Old Phrygian texts in a typical Phrygian alphabet dating from *c.* 730–450 BC, and the New Phrygian inscriptions (sepulchral texts in the Greek alphabet) stemming from the 1st and 2nd centuries AD. The Old Phrygian texts may be divided into a central group (Midas City and the central area) and an eastern group (found in Gordium), with offshoots in a still more eastern direction marking the utmost Phrygian expansion (inscriptions in or around Hüyük near Alaca, in Boğazköy-Hattusas, and in Tyana). An important recent finding—and the longest Old Phrygian text to date—is the rock inscription near the village of Germanos (modern Soğuk Çam) in Bithynia (found in 1966). The total number of Old Phrygian texts now stands at about 80; more than 50 of these are from Gordium and represent about one-quarter of the available material. There is a consensus of opinion on the Indo-European character of the Phrygian language; most scholars think that Phrygian is somehow connected to the Greek branch of Indo-European languages, although, at an earlier stage, some scholars considered the possibility of a connection with the Anatolian branch of Indo-European, and others proposed a relationship with Thracian and Illyrian.

Old Phrygian and New Phrygian

In a publication of new material from Gordium, the U.S. archaeologist Rodney S. Young cautiously suggested that the Old Phrygian alphabet may be dependent on a prototype in use on the North Syrian or Cilician coasts. The old idea that the Phrygian alphabet was dependent on a Greek one (and not vice versa) need not be abandoned in that case. Historically, such a derivation would present no problems, because the presence of Greek settlements in these areas during the second half of the 8th century BC is amply attested to by both Assyrian annals and late Greek historical sources as well as by archaeological findings. Internal evidence from the Phrygian alphabet, presented by the French linguist Michel Lejeune, serves as proof for some researchers that that alphabet derived from the Greek one.

Lydian. Of the more than 70 Lydian texts (*e.g.*, sepulchral inscriptions, votive texts, many graffiti), more than half have been found by United States excavators at the Lydian capital, Sardis. Two small Greek–Lydian bilingual texts were far less helpful than the famous Aramaic–Lydian text. A few texts (about ten) may go back to the 6th or 5th centuries BC, but many more stem from the 4th century. The Lydian alphabet was derived from an East Greek prototype; the superfluous signs in the Greek alphabet were used for specific Lydian sounds, and additional signs were either borrowed from other "Anatolian alphabets" or freely created.

Research on Lydian

Important results concerning Lydian were reached using a strictly combinatory method; *i.e.*, passages were compared that expressed similar contents in a slightly different manner in order to obtain a better understanding of the language's structure. This stage of the research culminated in a conclusive article by the Italian Piero Meriggi on the Indo-European character of Lydian (1936). Subsequently, other scholars published evaluations of the Lydian data, a dictionary, and a grammar book. The study of Lydian is hampered by many lexicological uncertainties, but there is at least a growing consensus on matters of grammar leading to the common notion that Lydian belongs to the Anatolian subgroup of Indo-European. The final obstacle to this classification as Anatolian was removed in 1959 by the Italian Onofrio Carruba, who proved that Lydian, like the other members of the Anatolian branch, does not possess a separate feminine gender. Lydian shares common features with Hittite, Palaic, and Luwian and should therefore be acknowledged, it seems, as a fourth independent member of the Anatolian subgroup.

Carian. A great number of the more than 100 Carian inscriptions are graffiti found in Egypt that were left behind by Carian mercenaries in the services of Egyptian pharaohs of the Saitic period (664–525 BC). In recent years, more monumental inscriptions have been found in Caria itself, and Carian clay tablets have also been discovered. In the mid-20th century, several scholars concluded that Carian writing consists of a purely alphabetic script and is not a mixed system of both single letters and syllabic signs as was formerly thought. It is a likely but still unproven assumption that Carian may also be classified in the Anatolian subgroup of Indo-European.

Lycian. More than 150 Lycian monumental inscriptions have been found so far, which, with very few exceptions, are sepulchral in character. They are written in an indigenous Lycian alphabet that is based not on an East Greek prototype (as its Lydian replica) but on a West Greek one. Although the Lycian coin legends are still usually dated from the period between 500 and about 360 BC, the tradition of the Lycian monumental inscriptions is now thought to have continued for a longer period, into the 3rd century BC. During the beginning of the research in the first half of the 19th century, extensive use was made of a good bilingual text that offers a faithful Greek translation. In the first phase of research, which ended about 1880, Lycian was investigated by an etymological method by which it was linked up either with Greek or the Iranian languages. A more reliable combinatory method was later introduced, but the most fecund phase in the study of Lycian occurred at the end of the 19th century, when the Scandinavian school of scholars cooperated closely in the publication of several important studies. In 1945, Holger Pedersen published a synthesis of all data that seemed to indicate a relationship of Lycian with Hittite; thus Pedersen proved conclusively that Lycian belongs to the Anatolian branch of Indo-European languages. This conclusion was slightly modified when the British scholar Franz J. Tritsch (in 1950), and, later, the French scholar Emmanuel Laroche showed that Lycian should be more specifically compared to Luwian.

Relationship of Lycian with Anatolian

Sidetic. The historical detail preserved by the Greek historian Arrian that the city of Side on the Pamphylian coast possessed a particular, indigenous language has been strikingly confirmed by legends on Sidetan coins of the 5th (?) through the 3rd (?) centuries and by five inscriptions from the 3rd and 2nd centuries BC (two of which are bilingual). There is a curious likelihood that this alphabet was directly derived from a Semitic writing system rather than from a Greek prototype, but Greek influence was not absent, as is clearly evidenced in the Greek bilingual texts and by a loanword from Greek. The first reliable study of Sidetic was made by Helmuth T. Bossert in 1950. In the case of Sidetic, even the value of a group of signs is still undecided, and research has not yet reached a stage in which a fruitful analysis of the texts and a classification of the language are within sight.

LINGUISTIC CHARACTERISTICS

Non-Indo-European languages. The non-Indo-European Hattic is an agglutinative language; that is, it combines several elements of meaning into a single word. In the conjugation of verbs, it uses prefixes that are attached to the word stems, which are mostly monosyllabic or bisyllabic. Hattic nouns consist of a free number of syllables and have both prefixes and suffixes. There are, however, no formal distinctive features to distinguish nouns and verbs.

Both the Hurrian and the Urartian languages differentiate between stems and suffixes, but there is again no sharp distinction between noun and verb. Many suffixes may be added onto one another in a row, but within the often prolonged suffix series a detailed order is rigidly observed. Among the suffixes added to the noun, several subgroups are distinguished; one group might be compared to the case endings of the Indo-European languages. One of the most characteristic phenomena of this group is the distinction between a subject case (the "nominative") and an "agentive" case. The agentive marks the actor or subject of a transitive verb when the object is expressed by its counterpart, the "nominative." The subject case is characterized by a lack of ending on the stem; it marks the subject in nominal sentences (sentences without verbs) and occurs with intransitive verbs and as the object of transitive verbs.

Subject and agentive cases in Hurrian and Urartian

Phrygian. The New Phrygian texts especially favour the attribution of the Phrygian language to Indo-European. They contain such data as *ios* as relative pronoun (Indo-European **io-s*, Greek *hos*), a demonstrative pronoun that is either comparable to Indo-European **ki-/ko-* or to **so* (an asterisk indicates a hypothetical reconstructed form), and the form *ad-daket* "he adds" related to Latin *addit* and to Greek *é-thē-ka.*

The Anatolian subgroup of Indo-European. *Grammatical characteristics.* Characteristic of the Anatolian languages is the absence of the dual number ("you and I") and the lack of feminine gender in the declension of nominals (nouns, pronouns, and adjectives). There is a division between an animate (common) gender and an inanimate (neuter) gender. In Hittite, a neuter may not be the subject of a transitive action verb; in that case, an *-ant* suffix is added before the neuter nominative ending in *-s.* This *-s* ending persists in the whole subgroup. The case system of Old Hittite is still fairly complicated, but in the subgroup as a whole there is a clear tendency toward a greater simplicity. The presence in Hittite of an archaic irregular class of nouns is a striking characteristic; *e.g.,* there are alternate *r* and *n* stems as in *uttar/uttanas* "word, affair" and *watar/witenas* "water."

The Anatolian inflection of pronouns conforms to the traditional Indo-European pattern by being different from that of the nouns, but, at the same time, it shows some striking peculiarities. Typical Anatolian pronouns are: Hieroglyphic Luwian *amu,* equivalent to Lycian *amu, emu, ẽmu* "I, me" (compare Hittite nominative *uk,* accusative *ammuk*); and Hittite nominative *zik,* accusative *tuk* "you," as compared to *ti, tu* in Palaic. Some of the languages have enclitic pronouns; *i.e.,* pronouns pronounced as being part of the preceding word. A demonstrative pronoun *aba-* ("that," but in some member languages also "this") is present in Hittite, Palaic, Cuneiform and Hieroglyphic Luwian, Lycian (*ebe-*), and Lydian (*ebad* "here, there"), and an interrogative or relative pronoun *kui-* (compare Latin *quis*) is common to Hittite, Palaic, and Cuneiform Luwian. The corresponding terms for *kui-* in Hieroglyphic Luwian, Lycian, and Lydian also seem to be phonetic variants of the same original pronoun.

The Anatolian verbal system is simple: it has two moods (indicative and imperative) and two tenses (present and preterite). There are some traces—either to be classified as debris or as the nucleus for a future development—of an aorist *-s* fixed to the stem; *e.g., kaness-* "to acknowledge" (compare Greek *gi-gnō-sk-ō*); *kalless-* "to call" (compare Greek *kaleō,* aorist *é-kale-s-a*). (The aorist is a verb form denoting action without reference to its duration or completion.) A mediopassive "voice" is present in Hittite (*es-a-ri* "he is seated"; *ki-tta-ri* "he is lying"), Palaic, Luwian, and perhaps in Lydian. (The mediopassive expresses a type of reflexive meaning ["He washes himself"] or passive meaning ["He is being washed"].)

Anatolian verbs

Reduplication (repetition) of the verbal stem occurs in the entire Anatolian subgroup. It adds an iterative or intensive nuance to the meaning, but it does not function in a system of tenses as in Greek. Very typical of the Anatolian subgroup are verbal suffixes like the causative *-nu-* (compare Hittite *war-* "to burn," *warnu-* "to kindle," *harg-* "to perish," *harganu-* "to ruin, to destroy"). In principle, these formations can be built on any verbal stem whose meaning permits such an addition. It should be stressed that in Hittite a normal expression for a "state" consists of a nominal sentence (that is, a sentence without a verb but with a noun, an adverbial expression, or a participle as predicate); sometimes, however, the verb *es-* "to be" is used as the carrier of modal or temporal nuances. The total absence of the Indo-European perfect (describing a "state" resulting from a recently concluded action) becomes very clear by the usage of the adverb *nawi* "not yet," which occurs with a present tense in Hittite (but which would employ a perfect tense, such as "has been," in English and other Indo-European tongues).

Very characteristic of the Anatolian subgroup is a strong preference for the linking together of particles and enclitic pronouns to form "chains" that are placed at the beginning of the sentence or clause. The first component of such a "chain" usually is a stressed part of the sentence or otherwise a sentence connective (like *nu* in Hittite, *a* in Luwian).

Phonological characteristics. In the Anatolian vowel system, *a, e, i,* and *u* are present, but *o* is curiously absent. In Lycian, the Greek value omicron has been used for the Lycian *u,* but in Lydian the existing *o* seems to be a secondary development. A main dialectal criterion is the treatment of Anatolian *e:* in Old Hittite, there still was a differentiation between *e* and *i,* but in later Hittite, an *-e* at the end of a word changed to *-i.* In Luwian, *e* tended to appear as *a.* Vowel gradation (*i.e.,* a change of vowel) that reflects meaning change plays a role in Hittite (*e-es-zi* "he is" versus *a-sa-an-zi* "they are") but was impossible in Luwian because of the sound change. Both Lycian and Lydian possess separate signs for nasalized vowels (*ã* and *ẽ*).

Advocates of the so-called laryngeal theory (first proposed by the Swiss linguist Ferdinand de Saussure in 1879) have found their postulate partly confirmed by Anatolian data. This theory maintains that the different forms of certain

Laryngeal theory and Anatolian data

Table 3: Anatolian Lexical Correspondences

English form	Hittite	Palaic	Cuneiform Luwian	Hieroglyphic Luwian	Lycian	Lydian
to make, to do	iya-		aya-	aia-	a-/e-	i- (?)
in	anda		anda	anta	ñta/e	(-)ẽn-/(-)ẽt-
to be	es-/as-	es-/as-	as-	as-	es-	e- (?)
house	pir-/parn-		parn-	parn-	*prñn-	bira-
to give	pai-/piya-	pi(ya)-	piya-	pi-	pije-	bi-
up	ser/sara		sarri	logogram	hri	
high, superior	sarazzi-			*sarli- (?)	hrzzi-	serli-
god	siu-/siwi-/ siwa-/siun-/ siuni-/siuna-		massani-	masana/i-	mahani-	civ-
divine	siunali-		massanalli-			civnali-

*Indicates that the following form is an unattested, reconstructed form.

words in the various Indo-European subgroups can be satisfactorily explained only by assuming that all the known Indo-European languages have lost certain guttural sounds (laryngeals) that were originally present in the parent speech. In 1927 the Polish linguist Jerzy Kuryłowicz and the French scholar Albert L.M. Cuny announced their discovery that in Hittite an *h* sound was preserved in positions in which a laryngeal would have formerly been (compare Hittite *hant-* "front" to English *anti-;* Hittite *pahhur* "fire" but English *pyre*). But the Anatolian evidence for the laryngeal theory is certainly not without problems, and the adherents of the theory consider that other laryngeals disappeared in Hittite as well.

Lexical data. Some examples of correspondences in vocabulary are given in Table 3. It has often been remarked—and not without reason—that although the grammar of the Anatolian languages would be recognizably Indo-European, the vocabulary would be less so. This is usually attributed to the deeply penetrating influences exercised by strange surroundings, not only during the period of time when the "Anatolians" were "en route" but also after their arrival in Anatolia.

The relationship with the other subgroups. The relationship of the Anatolian branch to the rest of Indo-European has often been defined in the United States on the basis of the "Indo-Hittite hypothesis"; that is, Hittite or Anatolian on the one hand and Proto-Indo-European on the other were both supposed to descend from a common parent. This hypothesis attributes too much weight to the Anatolian evidence. It was demonstrated as early as 1938 that the Anatolian branch should be placed on a par with the rest of the Indo-European subgroups and not as a coequal with Indo-European itself. The Indo-Hittite hypothesis is now rarely defended. Another extreme position states that the Hittite-Luwian-speaking group (another designation for the Anatolian subgroup) left the Indo-European parent group comparatively late, after the Greek and Armenian divisions had done so and approximately at the same time as Indo-Iranian. If this theory were true, there would be no need to use the Anatolian data for a thorough revision of the reconstructed Proto-Indo-European language, because these data would be less relevant, at least not more so than Indo-Iranian and Greek, on which the old reconstruction was based. A third opinion—prevalent in the French school of Indo-European studies—holds that the Hittite or, preferably, the "Common Anatolian" data are of special importance, because the Anatolian languages are particularly archaic. According to this theory, similarities in morphology (word elements) between the Celtic, Italic, and Hittite-Luwian groups and Tocharian (an Indo-European language of Central Asia) seem to imply that the dialects from which these groups evolved were in peripheral positions in the Indo-European language area and were probably the first to move away from the main group.

Indo-Hittite hypothesis

(Ph.H.J.H.t.C.)

Indo-Iranian languages

The Indo-Aryan languages and the Iranian languages together constitute the Indo-Iranian language group, the easternmost major branch of the Indo-European family of languages. Indo-Aryan (Indic) languages are spoken by some 800 million persons in India, Pakistan, Sri Lanka, Nepal, Bangladesh, and other areas of the Himalayan region. In addition, languages of the Indo-Aryan group are spoken by about 5,000,000 people in Europe, Africa, the Americas, and Oceania: the Gypsy, or Romany, dialects that are distributed about parts of Asia, the Middle East, Europe, and North America are of Indo-Aryan origin. Speakers of Iranian languages number in the tens of millions and live in areas extending from Pakistan to Iran, Afghanistan, Transcaucasia, and Central Asia. Among the Indo-European languages, only Linear A and Linear B and Hittite possess records that go back farther in time than those of Indo-Iranian.

The Indo-Iranian tongues have been used as both administrative and literary languages. Old Persian was the administrative language of the early Achaemenian dynasty dating from the 6th century BC; and an eastern Middle Indo-Aryan dialect was the language of the chancellery of the Mauryan emperor Aśoka in India in the mid-3rd century BC. As literary languages, the Indo-Iranian languages were used in the texts of some of the world's great religions: Indo-Aryan for Buddhism, Hinduism, and Jainism, and Iranian for Zoroastrian and Manichaean texts. The oldest Zoroastrian texts are in dialects included under the name Avestan. Commerce, conquest, and religion spread the influence of these languages. Indo-Aryan languages, for example, penetrated deep into Southeast Asia; names in Indonesia and other areas and Sanskrit texts in Cambodia reflect this influence.

The close relation between the Iranian and Indo-Aryan groups has never been doubted. They share characteristic features that set them apart as a subgroup of Indo-European. The long and short varieties of the Indo-European vowels *e, o,* and *a,* for example, appear as long and short *a:* Sanskrit *manas-* "mind, spirit," Avestan *manah-*, but Greek *mĕnos* "ardour, force." (In the following examples, a macron (¯) indicates a long vowel; a breve (˘) indicates a short vowel. The spellings used in this article for Indo-Aryan and Iranian forms are traditional transliterations for the most part. In some cases, more accurate phonetic symbols are used. These can be found in the International Phonetic Alphabet.) In instances in which some Indo-European languages have an *a* sound, Indo-Iranian has *i* as a reflex of Indo-European sounds called laryngeals—*e.g.*, Greek *patḗr* "father," Sanskrit *pitṛ-*, Avestan and Old Persian *pitar-*. After stems ending in long or short *a, i,* or *u,* an *n* occurs sometimes before the genitive (possessive) plural ending *ām* (Avestan *-ąm*)—*e.g.*, Sanskrit *martyānām* "of mortals, men" (from *martya-*); Avestan *mašyānąm* (from *mašya-*); Old Persian *martiyānām.*

Linguistic features shared by Indo-Aryan and Iranian

In addition to several other similarities in their grammatical systems, Indo-Aryan and Iranian have vocabulary items in common—*e.g.*, such religious terms as Sanskrit *yajña-*, Avestan *yasna-* "sacrifice"; and Sanskrit *hotṛ-*, Avestan *zaotar-* "a certain priest"; as well as names of divinities and mythological persons, such as Sanskrit *mitra-*, Avestan *miθra-* "Mithra." Indeed, speakers of both language subgroups used the same word to refer to themselves as a people: Sanskrit *ārya-*, Avestan *airya-*, Old Persian *ariya-* "Aryan."

The Indo-Aryan and Iranian language subgroups also differ from each other in a number of linguistic features, among them that Indo-Aryan has an *i* sound representing an Indo-European laryngeal sound not only in initial syllables but generally also in interior syllables; *e.g.*, Sanskrit *duhitṛ-* "daughter" (*cf.* Greek *thugátēr*). In Iranian, however, the sound is lost in this position; *e.g.*, Avestan *dugədar-*, *duγdar-*. Similarly, the word for "deep" is Sanskrit *gabhīra-* (with *ī* for *i*), but Avestan *ǰafra-*. Iranian also lost the accompanying aspiration (a puff of breath, written as *h*) that is retained in certain Indo-Aryan consonants; *e.g.*, Sanskrit *dhā* "set, make," *bhṛ*, "bear," *gharma-* "warm," but Avestan and Old Persian *dā*, *bar*, and Avestan *garəma-*. Further, Iranian changed stops such as *p* before consonants and *r* and *v* to spirants such as *f*: Sanskrit *pra* "forth," Avestan *frā*; Old Persian *fra*; Sanskrit *putra-* "son," Avestan *puθra-*, Old Persian *puṣsa-* (*ṣs* represents a sound that is also transliterated as *ç*). In addition, *h* replaced *s* in Iranian except before non-nasal stops (produced by releasing the breath through the mouth) and after *i*, *u*, *r*, *k*; *e.g.*, Avestan *hapta-* "seven," Sanskrit *sapta-*; Avestan *haurva-* "every, all, whole," Sanskrit *sarva-*. Iranian also has both *xš* and *š* sounds, resulting from different Indo-European *k* sounds followed by *s*-like sounds, but Indo-Aryan has only *kṣ*; *e.g.*, Avestan *xšayeiti* "has power, is capable," *šaēiti* "dwells," but Sanskrit *kṣayati*, *kṣeti*. Iranian was also relatively conservative in retaining diphthongs that were changed to simple vowels in Indo-Aryan.

Iranian differs from Indo-Aryan in grammatical features as well. The dative singular of *-a*-stems ends in *-āi* in Iranian; *e.g.*, Avestan *mašyāi*, Old Persian *cartanaiy* "to do" (an original dative singular form functioning as infinitive of the verb). In Sanskrit the ending is extended with *a*—*martyāy-a*. Avestan also retains the archaic pronoun forms *yūš*, *yūžəm* "you" (nominative plural); in Indo-Aryan the *-s-* was replaced by *y* (*yūyam*) on the model of the 1st person plural—*vayam* "we" (Avestan *vaēm*, Old Persian *vayam*). Finally, Iranian has a 3rd person pronoun *di* (accusative *dim*) that has no counterpart in Indo-Aryan but has one in Baltic.

Original homeland of Indo-Iranian group

The original location of the Indo-Iranian group was probably to the north of modern Afghanistan, in the present-day states of Tajikistan, Uzbekistan, Kyrgyzstan, Turkmenistan, and Kazakhstan, where Iranian languages are still spoken. From there, some Iranians migrated to the south and west, the Indo-Aryans to the south and east. From geographical references in the earliest Indo-Aryan literary document, the Rigveda, it is clear that the earliest settlement of Indo-Aryans was in the northwest of the Indian subcontinent. Migration did not take place at once; there was doubtless a series of migrations. The date of entry of the Indo-Aryans into the subcontinent cannot be precisely determined, though the beginning of the 2nd millennium BC is plausible and generally accepted.

There is heated controversy concerning the precise linguistic position of the language of the Indo-Iranian family first attested in Middle Eastern cuneiform texts of *c.* 1450–1350 BC. Some borrowed words and proper names appearing in these Hittite-Hurrian documents have been interpreted as belonging either to Indo-Iranian, to an Indic subgroup of Indo-Iranian that had not yet fully split, or to Indo-Aryan proper. Complete scholarly agreement on this issue has not been reached.

The identification of the Harappan peoples of the Indus Valley, whose writing has not yet been satisfactorily deciphered, also awaits further research; with it may come a possible answer as to whether Indo-Aryans encountered

Figure 3: Distribution of the Indo-Iranian languages.

these people or whether their civilization had passed by the time the Indo-Aryans arrived on the subcontinent. Whatever the answers to these problems may be, the reasons for the split of the Indo-Aryans and Iranians are not known.

In the following presentation regarding Indo-Aryan documents as evidence for linguistic history, it should be borne in mind that almost all dates are approximations.

THE INDO-ARYAN LANGUAGES

Languages of the group. Indo-Aryan languages are assigned to three major periods: Old, Middle, and New Indo-Aryan. These periods are linguistic, not strictly chronological. Old Indo-Aryan includes different dialects and linguistic states referred to in common as Sanskrit. The most archaic Old Indo-Aryan is that of sacred texts called Vedas. Classical Sanskrit is the name given to the literary language that represents a polished form of various dialects. The late Vedic dialect described by the grammarian Pāṇini (*c.* 6th century BC) is also commonly called Classical Sanskrit. Middle Indo-Aryan includes both the dialects of inscriptions from the 3rd century BC to the 4th century AD and literary languages. Apabhraṃśa dialects represent the latest stage of Middle Indo-Aryan development. Though all Middle Indo-Aryan languages are included under the name Prākrit, it is customary to speak of the Prākrits as excluding Apabhraṃśa.

New Indo-Aryan is represented by such modern vernaculars as Hindi and Bengali, which began to emerge from about the 10th century AD. These too have earlier and later stages, culminating in the present-day languages.

Modern Indo-Aryan tongues and number of speakers

New Indo-Aryan languages accounted for about 490,000,000 speakers in India, or approximately 74 percent of the population in the early 1980s. Considering the approximately 85,000,000 Bengali speakers in Bangladesh, approximately 63,000,000 speakers accounted for by Punjabi and Sindhi in Pakistan, and 11,000,000 Sinhalese (Sinhala) speakers in Sri Lanka (formerly Ceylon), the total number of New Indo-Aryan speakers is well over

Table 4: Modern (New) Indo-Aryan Languages
key: B—Bangladesh, former East Pakistan; I—India; N—Nepal; P—Pakistan (former West Pakistan)

language group, language	where principally spoken*	estimated number of speakers, 1981 (000)†	comments
Eastern group			
Assamese	*Assam*, India	11,800 (I)	official language of Assam, India; also‡
Bengali	*Bangladesh; West Bengal, Tripura*, and Assam, India	85,300 (B); 58,500 (I); 90 (P); 30 (N)	official language of Bangladesh and of West Bengal, Tripura, and Manipur, India; also‡
Oriya	*Orissa*, India	24,700 (I); 23 (B)	official language of Orissa, India; also‡
Northwest group			
Punjabi	*Punjab*, Northwest Frontier Province and Karāchi, Pakistan; *Punjab*, Haryana, Delhi, and Ganganagar district of Rajasthān, India; *Jammu* portion of both Indian- and Pakistani-held portions of Jammu and Kashmir	52,800 (P)§; 17,100 (I); statistics not available for Pakistani-held portion of Jammu and Kashmir	official language of Punjab, Pakistan, and with Urdu, of Jammu section of Jammu and Kashmir
Lahnda	Punjab and Northwest Frontier Province, Pakistan	statistics not available (P)¶, 14 (I)	
Sindhi	*Sind* province and *Las Bela* and other eastern districts of Baluchistan province, Pakistan; Kutch district of Gujarāt, India	10,100 (P); 2,040 (I); mainly immigrants from area now in Pakistan	official language of Sind, Pakistan; also‡
Pahari (a group of languages)			
Eastern Pahari			
Nepal (major east Pahari language)	*Nepal, Sikkim*, Bhutan	7,630 (N); 1,600 (I); 74 (Sikkim 1961), statistics not available for Bhutan	official language of Nepal
Central Pahari			
Kumauni	Himalayan Uttar Pradesh, India	1,500 (I)	
Garhwali	Himalayan Uttar Pradesh, India	1,160 (I)	
Western Pahari (62 languages and dialects according to the Indian census of 1961)	*Himachal Pradesh*, adjoining district of Uttar Pradesh, India; Himalayan districts of Indian- and Pakistani-held Jammu and Kashmir	994 (I); statistics not available for Pakistani-held portion	Pahari is an official language of Himachal Pradesh, along with Hindi
Unclassified and unspecified	(same as Western Pahari, above)	1,500 (I)	
Dardic			
Dard (East Dardic)			
Kashmiri (major language of East Dardic)	*Vale of Kashmir* and adjoining districts to south and west in Indian- and Pakistani-held portions of Jammu and Kashmir	2,995 (I), of which 2,991 are speakers of Kashmiri; 74 (P); statistics not available for Pakistani-held portion of Jammu and Kashmir or for Afghanistan	Kashmiri, along with Urdu, is an official language of the Kashmiri-speaking area of Jammu and Kashmir; also‡
Other Dardic languages:			
Khowari (Central Dardic), Kafiri (West Dardic), and other minor languages	*Gilgit Agency* of Pakistani-held portion of Jammu and Kashmir; adjoining districts of Northwest Frontier Province, Pakistan; and adjoining portion of northeast Afghanistan	208 (P, partial enumeration); statistics not available for Pakistani-held portion of Jammu and Kashmir or for Afghanistan	position of Dardic is disputed; some account for its peculiarities by proposing that it left the Indo-Iranian branch after Indo-Aryan but before all the features particular to Iranian had evolved; others suggest that East and Central Dardic are definitely Indo-Aryan, but that they did not go through the middle Indo-Aryan stage represented in documents; Kafiri occupies a special portion
West and Southwest groups			
Gujarati	*Gujarāt*, Bombay district of Mahārāshtra, India	32,330 (I); 503 (P);	official language of Gujarāt, India; also‡
Marathi	*Mahārāshtra* and eight adjoining districts in three older states of India	51,800§ (I)	official language of Mahārāshtra, India; also‡
Konkani	*Goa*, coastal Mahārāshtra south of Bombay, and coastal Karnataka, India	1,700¶ (I)	
Sinhalese	*Sri Lanka* (Ceylon)	10,986 (census 1981 data of Sri Lanka for Sinhalese group)	official language of Sri Lanka (Ceylon)
Divehi (Maldivian)	*Maldive Islands*	approximately 150	official language of Maldive Islands

*Italic type indicates language is spoken by a majority or plurality of the population in the area; roman type indicates language is spoken by a minority of the population in the area. †Not shown when there are fewer than 10,000 reported speakers in a given country. "Reported" number of speakers is often far different from actual number of speakers. Indian (I) data include census returns for Indian-held portion of Jammu and Kashmir; Pakistani (P) data do not include data for Pakistani-held portion of Jammu and Kashmir; Pakistani data are incomplete for the tribal areas of the Northwest Frontier Province, west Dardic

650,000,000. According to the latest Indian census, there are 547 mother tongues of the Indo-Aryan group in use within the bounds of postpartition (1947) India. Some of these are dialects that are used by few speakers; others are official state languages having 30,000,000 or 50,000,000 speakers. The major groups of New Indo-Aryan languages are given in Table 4. Structurally and historically, Hindi and Urdu are one, although they are now official languages of different countries written in different alphabets. The term *hindī* (also *hindvī*) is known from as early as the 13th century. The term *zabān-e-urdū* "language of the imperial camp" came into use in about the 17th century. In the south, Urdu was used by Muslim conquerors of the 14th century.

Official state languages

Many of the languages in Table 4 are official state languages, the media of education up to the university level and of official transactions. Hindi, written in the Devanāgarī script, is the co-official language (with English) of the Republic of India and is used as a lingua franca throughout North India. It has varieties according to the mother tongue of the area; *e.g.*, Bombay Hindi and Calcutta Hindi. Each of the major state languages has several other dialects in addition to the standard dialect adopted for official purposes. Including the various dialects down to the village level, it can be said that a chain of communication stretches across North India such that each dialect forms a link with each adjacent dialect. On the level of official languages this is not so: a Gujarati speaker will not readily understand colloquial Bengali.

Historical survey of the Indo-Aryan languages. The points noted above regarding Indo-Aryan migration make it difficult to determine the domain of Proto-Indo-Aryan, the ancestral language of all the known Indo-Aryan tongues, if indeed there was any such single region. All

Table 4: Modern (New) Indo-Aryan Languages (continued)

language group, language	where principally spoken*	estimated number of speakers, 1981 (000)†	comments
Midland group			
Hindi	*Uttar Pradesh, Madhya Pradesh, Bihār, Haryana, Delhi,* Rājasthān, Punjab, Himachal Pradesh, and in scattered proximate districts of West Bengal and Mahārāshtra, India	173,000§ (I); 249 (B); 225 (N)	co-official language (with English) of the Republic of India and a lingua franca throughout North India; language of official business in area described; Khari Boli, based on a dialect of western Uttar Pradesh to the northeast of Delhi, is considered to be a standard form of Hindi; official language of Uttar Pradesh, Madhya Pradesh, Bihār, Haryana, Rājasthān, and Himachal Pradesh states and of Delhi union territory, India, also‡
Eastern Hindi (incomplete)			
Awadhi (Avadhi)	north central and central Uttar Pradesh, India	153 ‖ (I)	
Bagheli	north central Madhya Pradesh and south central Uttar Pradesh, India	421 ‖ (I)	
Chattisgarhi	east central Madhya Pradesh, India	7,495§ (I)	
Western Hindi (incomplete)			
Braj Bhasa	western Uttar Pradesh and adjacent districts of Haryana, Rajasthan, and Madhya Pradesh, India	32 ‖ (I)	
Bundeli (Bundelkhandi)	north central Madhya Pradesh and southwestern Uttar Pradesh, India	470 ‖ (I)¶	
Other Hindi languages and dialects (Eastern and Western, including Hindustani)	scattered over much of Uttar Pradesh, Madhya Pradesh, and Haryana and in eastern Rājasthān, India	10,066 ‖ (I)	
Urdu	*Karāchi* district and Pakistan in general; all but northeastern and southern peninsular India	35,000¶ (I); 6,700 (P), claimed as an additional language by 5,900 others (P); 240 (B); 54 (N)	official language of Pakistan (before 1971 co-official language of Pakistan; recognized in the constitution of India; a form of Urdu known as Dakhini Urdu (Southern Urdu) is still used in the area around Hyderābād; an official language in both the Indian- and Pakistani-held portions of Jammu and Kashmir; also‡
Bihari (a group of languages)			
Maithili	North Bihār, India; adjacent lowland Nepal	7,500 (I)¶; 1,560 (N)	
Magahi (Magadhi)	central Bihār, India	8,130 (I)	
Bhojpuri	western Bihār and eastern Uttar Pradesh, India	17,570 (I)¶; 1,120 (N)	
Others	Bihār, India	1,560 (I)¶	
Rajasthani (a group of languages			
Mewati	northeast Rājasthān, India	125 ‖ (I)	
Ahirwati	northeast Rājasthān	37 ‖ (I)	
Harauti	southeast Rājasthān	979 ‖ (I)	
Malvi	western Madhya Pradesh and southeast Rājasthān	1,190 ‖ (I)	
Nimadi	southwest Madhya Pradesh	920 ‖ (I) India	
Marwari	western, central, and northern Rājasthān	10,887¶ (I)	
Rajasthani-other and unclassified	over most of Rājasthān, with locally important groups in scattered districts of Mahārāshtra, Andhra Pradesh, and Karnataka, India; and both Indian- and Pakistani-held portions of Jammu and Kashmir	11,140¶ (I); 335 (P); statistics not available for Pakistani-held portion of Jammu and Kashmir	Rājasthāni, an official language of Rājasthān along with Hindi
Bhili (a group of dialects)	southern Rājasthān, western Madhya Pradesh, eastern Gujarāt, and northwest Mahārāshtra, India	1,562 ‖ (I)	
Khandeshi	northwest Mahārāshtra	750 (I)	
Others (inadequately classified or unspecified)			
Tharu	sub-Himalayan Nepal	603 (N)	
Miscellaneous dialects	sub-Himalayan Nepal	891–953 (N), depending on inclusiveness of listing	

languages, and Punjabi. As of 1972, Afghanistan has had no population census, but speakers of various Dardic languages there may number as many as 100,000, while another 10,000–20,000 may speak Indo-Aryan languages. ‡One of the 15 official languages listed in Schedule VIII of the Indian Constitution. §Presumed significant overstatement by census. Punjabi in Pakistan includes Lahnda. Marathi in India includes many speakers of Konkani and possibly also of Khandeshi. "Hindi" (undifferentiated) includes many speakers of all the languages of the Midland group footnoted ‖ or ¶. ‖ Presumed gross understatement by census. ¶Presumed significant understatement by census.

that can be said with certainty is that the Indo-Aryans on the subcontinent first occupied the area comprising most of present-day Punjab (both West and East), Haryana, and the Upper Doab (Ganges–Yamuna interfluve) of Uttar Pradesh. The structure of Proto-Indo-Aryan must have been close to that of early Vedic, with dialectal variations.

Vedic documents

Old Indo-Aryan. The most archaic Sanskrit is that of the Vedas, of which there are four major text groups called Saṃhitās: the Rigveda, Atharvaveda, Sāmaveda, and Yajurveda. The Yajurveda is in turn divided into two main branches, the White (Śukla) Yajurveda and the Black (Krishna) Yajurveda. The Rigveda, Atharvaveda, and Sāmaveda are purely metrical texts mainly used by priests in their ritual. The texts of the Black Yajurveda contain both verses used in ritual sacrifice (called *mantras*) and prose sections that are explanatory in nature, giving mythological explanations of sacrifices and objects used in them, together with etymologies (derivations of words). These sections are known as *Brāhmaṇa* portions. Each Veda also has a particular *Brāhmaṇa* connected with it. The early Vedic texts are pre-Buddhistic; a plausible date accepted for the composition of the Rigveda is between 1200 and 1000 BC, though the exact chronology of these early texts is difficult to establish. The prose passages of *Brāhmaṇas* and of the early *sūtra* (aphoristic texts) period may be called late Vedic. Also of the late Vedic period is the grammarian Pāṇini, author of a treatise called *Aṣṭādhyāyī*, who makes a distinction between the language of sacred texts (*chandas*) and the usual language of communication (*bhāṣā*).

Epic Sanskrit is so called because it is represented principally in the two epics, *Mahābhārata* and *Rāmāyaṇa*. In the latter the term *saṃskṛta* "formed, polished" is encountered, probably for the first time with reference to the language. The date of com-

Table 5: Sanskrit (Devanāgarī Alphabet and Numerals)

vowels and diphthongs: initial	medial	name	equivalents: EB preferred	alternatives	approximate* pronunciation
अ		akāra	a		f*u*n
आ	ा	ākāra	ā		f*a*ther
इ	ि	ikāra	i		f*i*ll
ई	ी	īkāra	ī		mach*i*ne
उ	ु	ukāra	u		p*u*ll
ऊ	ू	ūkāra	ū		r*u*de
ऋ	ृ	ṛkāra	ṛ	ṛi, *ri*	litt*er*
ॠ	ॄ	ṝkāra	ṝ	ṝi, *ri*	†
ऌ	ॢ	ḷkāra	ḷ	ḷri, *li*	ab*le*
ए	े	ekāra	e	ē	f*a*de
ऐ	ै	aikāra	ai	āi	s*i*te
ओ	ो	okāra	o	ō	b*o*ne
औ	ौ	aukāra	au	āu	n*ow*

consonants and special signs	name	equivalents: EB preferred	alternatives	approximate* pronunciation
Gutturals‡				
क	kakāra	k		*k*in
ख	khakāra	kh		bloc*kh*ead
ग	gakāra	g		*g*o
घ	ghakāra	gh		lo*g h*ut
ङ	ṅakāra	ṅ	ñ	si*ng*
Palatals				
च	cakāra	c	ch, *k*	*ch*in
छ	chakāra	ch	chh, *kh*	pi*tch h*ook
ज	jakāra	j	*g*	*j*ob
झ	jhakāra	jh	*gh*	he*dgeh*og
ञ	ñakāra	ñ	n	ca*ny*on
Retroflexed∯				
ट	ṭakāra	ṭ	*t*	po*t*
ठ	ṭhakāra	ṭh	*th*	an*th*ill
ड	ḍakāra	ḍ	*d*	di*d*
ळ	ḷakāra	ḷ	*l*	‖
ढ	ḍhakāra	ḍh	*dh*	a*dh*ere
ण	ṇakāra	ṇ	*n*	ow*n*
Dentals¶				
त	takāra	t		li*t*tle
थ	thakāra	th		boa*t h*ouse
द	dakāra	d		*th*en
ध	dhakāra	dh		an*d h*e
न	nakāra	n		*n*o
Labials♀				
प	pakāra	p		li*p*
फ	phakāra	ph		u*ph*ill
ब	bakāra	b		*b*aby
भ	bhakāra	bh		a*bh*or
म	makāra	m		*m*aim
Semi-vowels				
य	yakāra	y		*y*ard
र	repha	r		*r*are
ल	lakāra	l		*l*ily
व	vakāra	v		*w*e
Spirants♂				
श	śakāra	ś	*ç, s*	*sh*y (palatalized)
ष	ṣakāra	ṣ	sh	*sh*y (retroflexed)
स	sakāra	s		*s*and
ह	hakāra	h		*h*at
Diacritics				
ः	visarga	ḥ		□
ं	anusvāra	ṃ	n̲	◇
ँ	anunāsika	ṃ	ṁ	◇

numerals

Devanāgarī	Arabic	Devanāgarī	Arabic	Devanāgarī	Arabic
०	0	११	11	२२	22
१	1	१२	12	२३	23
२	2	१३	13	२४	24
३	3	१४	14	२५	25
४	4	१५	15	२६	26
५	5	१६	16	२७	27
६	6	१७	17	२८	28
७	7	१८	18	२९	29
८	8	१९	19	३०	30
९	9	२०	20	१००	100
१०	10	२१	21	१०००	1,000

*These pronunciations apply to Sanskrit. The same symbols sometimes have different values in the modern languages. †Same as *ṛkāra*, but lengthened. ‡Pronounced at back of throat. ∯Pronounced with the tongue curled back against the roof of the mouth. ‖A retroflexed *l*, close to the second *l* in "little." ¶Pronounced with closed teeth. ♀Pronounced with the lips. ♂Breathed. □A diacritical mark indicating aspiration. ◇Diacritical marks indicating nasalization.

position for the core of early Epic Sanskrit is considered to be in the centuries just preceding the Christian era.

Classical Sanskrit is the language of the major poetic works (*kāvya*), drama (*nāṭaka*), tales such as the *Hitopadeśa* and *Pañca-tantra*, and technical treatises on grammar, philosophy, and ritual. It was used not only by the poet Kālidāsa and his predecessors Bhāsa, a dramatist, and Aśvaghoṣa, a Buddhist author, in the first centuries AD but was also continued long after Sanskrit was a commonly used mother tongue; indeed, Sanskrit is a language of learned treatises and commentaries to this day. It is also used as a lingua franca among *paṇḍit*s (Brahmin scholars) from different areas of India.

Development of Sanskrit

Linguistic developments can be traced from the early Vedic of the Rigveda through the later Saṃhitās on to the late Vedic of *Brāhmaṇa* prose and *sūtra*s, culminating in the language described by Pāṇini, which is tantamount to Classical Sanskrit. For example, the nominative plural form ending in *-āsas* (*devāsas* "gods") was already less frequent than *-ās* in the Rigveda and continued to lose ground later; in *Brāhmaṇa, -ās* (*e.g., devās*) is the normal form. There are numerous other changes evident. For example, the instrumental singular form of *-a*-stems ends both in *-ā* and *-ena* (a pronoun ending) in the Rigveda, with the latter form predominating; thus, *vīryā* "heroic might" appears once, and *vīryeṇa* occurs ten times (from *vīrya-* "heroic might, act"). In later Vedic *-ena* is the usual ending. All the early Vedic forms are expressly classed as belonging to the sacred language (*chandas*) by Pāṇini.

The verb also shows chronological differences. For example, the 1st person plural ending *-masi* (*e.g., bharāmasi* "we bear") predominates over *-mas* in Rigvedic but not in the Atharvaveda; *-mas* becomes the normal ending later. Early Vedic distinguishes between the aorist, imperfect, and perfect tenses. The aorist is commonly used to refer to an action that has recently taken place; the imperfect is a narrative tense referring to actions accomplished in the distant past. The perfect form of the verb originally denoted, as in Greek, a state reached; *e.g., bi-bhāy-a* "is afraid" (root *bhī*). From earliest Vedic, however, this was not always the use of the perfect. Although the grammarian Pāṇini distinguished between the three tenses noted (he said the perfect is used to denote an action beyond one's ken), the perfect and imperfect both came to be used as narrative tenses.

There are also future forms of Vedic, formed with suffixes (*-iṣya* and *-sya*) and used from earliest times. A future form, composed of an agent noun of the type *kartṛ-* "doer" and followed, except in the 3rd person, by forms of the verb *as* "be" (*e.g., kartāsmi* [*kartā asmi*] "I will do"), was recognized as in common use by Pāṇini but is rare in early Vedic.

The injunctive of early Vedic

Early Vedic had a category that went out of use by the late Vedic period of *Brāhmaṇas*—the injunctive, which was formally a form with secondary endings lacking the augment, a prefixed vowel. The injunctive could be used to denote a general truth. A general truth can also be signified by the subjunctive, which is characterized by the vowel *a* affixed to the present, aorist, or perfect stem. Later Vedic retained the injunctive only in negative commands of the type *mā vadhīs* "do not slay." The subjunctive also diminished slowly until it was no longer used; for Pāṇini the subjunctive belonged to sacred literature. The functions of the subjunctive were taken over by the form called optative (and the future form).

Noun forms incorporated into the verb system are numerous in early Vedic. Rigvedic has forms with affixes *yà* and *tva* functioning as future passive participles (gerundives); *e.g., vāc-ya-* "to be said," *kar-tva-* "to be performed, done." The Atharvaveda has, additionally, forms with *-(i)tavya* (*hiṃs-itavya-* "to be injured") and *-anīya* (*upa-jīv-anīya-* "to be subsisted upon"). By late Vedic, the type with *tva* had been eliminated; Pāṇini recognized as normal the types *kārya-, kartavya-, karaṇīya-* "to be done." In Indo-Aryan, from earliest Vedic down to New Indo-Aryan, forms called absolutives (or gerunds) are used to denote the previous of two or more actions performed (usually) by one agent: "having done . . . he did"; for example, *pibā niṣadya* "sit down (*niṣadya* "having sat down") and drink." Rigvedic uses *tví, tvā, tvāya, (t)ya* to form absolutives, but these were later reduced to two: *tvā* with a simple verb or one compounded with the negative particle, and *ya* with a verb compounded with a preverb (a preposition-like form).

Early Vedic also uses various case forms of action nouns in the capacity of infinitives; *e.g.*, dative singular *-tave* (*dā-tave* "to give"), genitive singular *-tos* (*dā-tos*), both from a noun in *-tu*, which also supplies the accusative ending *-tum* (*dā-tum*). There are other types in early Vedic, but the nouns in *-tu* are important; in late Vedic the accusative *-tum* and the genitive *-tos* (construed with *īś* or *śak* "be able, can") became the norm. According to Pāṇini, forms in *-tum* and dative singular forms of action nouns are equivalent variants: *bhok-tuṃ gacchati/ bhojanāya gacchati* "He is going out to eat."

Chronological and dialectical modifications

That some forms fell into disuse in the course of Indo-Aryan is natural; the above represent both chronological and dialectal modifications. Such change was recognized by Indian grammarians; *e.g.*, Patañjali, of the mid-2nd century BC, noted that perfect forms of the type *ca-kr-a* "you did, have done" (2nd person plural) were not in use at his time; instead, a nominal (adjective) form *kṛ-ta-vant-as* was used, consisting of the past passive participle *kṛ-ta-* and an adjectival suffix *-vant*. Indian grammarians also recognized the existence of different dialects. Pāṇini noted forms used by northerners (*udīcya*) and easterners (*prācya*), as well as various dialectal uses described by grammarians who preceded him.

Earlier documents also afford evidence for dialect variation; *e.g.*, the early Vedic of the Rigveda is a dialect in which the Indo-European *l* sound was for the most part replaced by *r*—*prā* "fill," *pūr-ṇa-* "full." This change accords with Iranian; *e.g.*, Avestan *pərəna* "full." These forms contrast with Latin *plenus* and Gothic *fulls*, with *l*. Other dialects kept *l* and *r* distinct. There are also doublets that have both *r* and *l* in words with Indo-European *r*: *rohita-/lohita-* "red." The variant with *l* can be assumed to belong to an eastern dialect. This variance accords with Middle Indo-Aryan evidence and the fact that such *l* forms become more numerous in the tenth book (*maṇḍala*) of the Rigveda, which is demonstrably more recent than the most ancient parts of the Rigveda and dates from a time when the Indo-Aryans had progressed farther east than their original location on the subcontinent. The development of retroflex *ḷ-* and *ḷh-* sounds (produced by curling the tip of the tongue upward toward the hard palate) from the retroflex sounds of *ḍ* (*nīḷa-* "nest" from *nīḍa-*) and *ḍh* when occurring between vowels is another feature characteristic of some dialects, including the major dialect of the Rigveda.

Classical Sanskrit and its accentual system

Classical Sanskrit represents a development of one or more such early Old Indo-Aryan dialects. At this state, the archaisms noted above have been eliminated. Moreover, the accentual system of Classical Sanskrit is not the same as that of Vedic, which had a system of pitches; vowels had low, high, or circumflex (first rising, then falling) pitch, and the particular vowel of a word that received high pitch could not be predicted. In Classical Sanskrit, on the other hand, the accent was probably predictable. If the next to the last vowel was long, it received the accent; if not, the vowel preceding it was accented. The Vedic system survived at least to the time of Pāṇini, who described it fully and did not restrict it to sacred language.

For all this simplification, Classical Sanskrit is considerably more complex than Middle Indo-Aryan. In addition to the vowels *a, i,* and *u* (in both long and short varieties), it has *ṛ* and *ḷ* used as vowels. Consonant clusters occur freely, except in word final position, and the system of sound modification conditioned by the context, called *sandhi*, is fully operative. Moreover, in its grammatical system Classical Sanskrit maintains the dual number, seven cases in addition to the vocative form (which marks the one addressed), and a complex set of alternations. For example, to the nominative singular form *agni-s* "fire," correspond the genitive singular *agne-s* "of fire" the nominative plural *agnay-as* "fires," and the instrumental plural *agni-bhis* "with fires," with differing vowels in the second syllable. There are also separate sets of nominal (noun) and pronominal (pronoun) endings. Some nouns and adjectives inflect as pronouns; *e.g.*, *ekasmai*, dative singular masculine-neuter of *eka-* "one."

The verb system of Classical Sanskrit also maintains complex alternations. In the present tense of the type *bhav-a-ti* "becomes, is," the stem (*bhav-a-*) remains unchanged throughout the paradigm except for lengthening of the *-a-* to *-ā-* before *v* and *m*. But other verbs have vowel alternation; *e.g.*, *as-mi* "I am," *s-mas* "we are"; *e-mi* "I go," *i-mas* "we go"; *juhomi* "I pour," *juhumas* "we pour." A distinction is observed between active and mediopassive endings: *jan-ay-a-ti* "engenders" with the active ending *-ti*, but *jā-ya-te* "is born" with the mediopassive ending *-te*. (Mediopassive verb forms are used for the passive, reflexive, and other meanings.)

Classical Sanskrit also has a rich system of nominal and verbal derivatives. Compound words are of the following kinds: copulative (*dvandva*) compounds such as *mātāpitarau* "mother and father" (also elliptic *pitarau* "parents"); the type like *tat-puruṣa-* "his man," in which the first member is equivalent to a case other than nominative; the type like *bahu-vrīhi* "much-rice," in which the object denoted is other than that of any of the members of the compound (*bahur vrīhir yasya* "He who has much rice"); and adverbial compounds (*avyayībhāva*) of the type *upāgni* (*upa-agni*) "near the fire." In addition, there are derivatives with affixes *-tara-* and *-tama*, such as *priya-tara-* "very dear" and *priya-tama-* "most dear" from the adjective *priya-*. Pronouns have derivatives equivalent to case forms; *e.g.*, *tatra* "there," *yatra* "where," and *kutra* "where?" are equivalent to locative forms such as *tasmin*, *yasmin*, and *kasmin*. These can also be used without a noun.

Among the derivative verbal systems are the causative and the desiderative ("desire to"); the former has an affix *-ay-* (*gam-ay-a-ti* "makes to go," *kār-ay-a-ti* "has do") or, after roots in *-a*, *-pay-* (*sthā-pay-a-ti* "sets in place"). The desiderative is formed with *-sa-* and reduplication (repetition of a part of the root)—*dī-dṛk-ṣa-te* "desires to see" (root *dṛś*). The desiderative also has an agent noun in *-u*—*dī-dṛk-ṣ-u* "who wishes to see."

The Prākrits

Middle Indo-Aryan. The Sanskrit word *prākṛta*, whence the term Prākrit, is a derivative from *prakṛti-* "original, nature." Grammarians of the Prākrits generally consider the original from which they derive to be the Sanskrit language as described by grammarians going back to Pāṇini. Most modern scholars consider *prākṛta* to refer to the "natural" languages, the vernaculars, as opposed to Sanskrit, the polished language of literature and the educated (*śiṣṭa*). There is also linguistic evidence to support this view. Several forms in the Prākrits are found in Vedic but not in Classical Sanskrit. As Classical Sanskrit is not directly derivable from any single Vedic dialect, so the Prākrits cannot be said to derive directly from Classical Sanskrit.

The most archaic literary Prākrit is Pāli, the language of the Buddhist canon (*c.* 5th century BC) and of the later stories and commentaries of Theravāda Buddhism. Pāli represents essentially a western Middle Indo-Aryan dialect, though there are sufficient easternisms in the canon to have led some scholars to the view that the canon as it exists today is a recast of an original in an eastern dialect. To the Buddhist literature also belongs the *Gāndhārī Dhammapada*, the only literary text written in a dialect of the northwest. The Niya documents, official documents written in Prākrit dating from the 3rd century AD, also belong to the northwest. The earliest inscriptional Middle Indo-Aryan is that of the Aśokan inscriptions (3rd century BC). These are more or less full translations from original edicts issued in the language of the east (from the capital Pāṭaliputra in Magadha, modern Patna in Bihār) into the languages of the areas of Aśoka's kingdom. There are other Prākrit inscriptions up to the 4th century AD, and Sanskrit was not used inscriptionally until the first centuries AD. Literary Prākrits other than Pāli were also used in independent works and in dramas along with Sanskrit.

According to Prākrit grammarians, Mahārāṣṭrī ("From the Mahārāshtra Country") is the Prākrit par excellence. It is the language of *kāvya*s (epic poems) such as the *Rāvaṇavaha* (also called *Setubandha*) from no later than the 6th century AD. Mahārāṣṭrī is also the language of lyrics in Rājaśekhara's *Karpūra-mañjarī* (*c.* 900), the only extant drama written completely in Prākrit, and of verses recited by women in the classical drama of Kālidāsa and his successors, though not earlier. The literary dialect used for conversation among higher personages other than the king and his captains in the drama is Śaurasenī, while Māgadhī is used by lower personages.

The language of the early Jaina canon, the final version of which was made in the 5th or 6th century AD, is called Ardhamāgadhī ("Half Māgadhī"); Jaina also used another literary dialect, called Jaina Māhārāṣṭrī in non-canonical works. The oldest poetic work in this is Vimala Sūri's *Paumacariya* (*c.* 3rd century). Of other Prākrit dialects mentioned by grammarians, Paiśācī (or Bhūta-Bhāṣā, both meaning "Language of Demons") is noteworthy; it is said to be the language of the original *Bṛhatkathā* of Guṇāḍhya, source of the Sanskrit book of stories *Kathāsaritsāgara*.

Buddhist Hybrid Sanskrit

Buddhist works were also written using a language that has been called Buddhist Hybrid Sanskrit. Among these works is the *Mahāvastu*, the core of which is thought to date from the 2nd century BC. This language is a Middle Indo-Aryan dialect of indeterminate origin, which steadily became more Sanskritized in prose sections of later works.

The most advanced stage of Middle Indo-Aryan, Apabhraṃśa, was also used as a literary language. That there was literary creation in Apabhraṃśa by the 6th century is clear from an inscription of King Dharasena II of Valabhī, in which the King praises his father as being adept in Sanskrit, Prākrit, and Apabhraṃśa composition. Moreover, in the fourth act of Kālidāsa's drama *Vikramorvaśīya* there are Apabhraṃśa verses. Because Kālidāsa probably lived in the 3rd or 4th century, literary composition in Apabhraṃśa is earlier still, if these verses are legitimate. There is a great deal of later literature in Apabhraṃśa, for the most part Jaina works; *e.g.*, *Paumacariu* of Svayambhū (8th–9th century), *Harivaṃśa-purāṇa* of Puṣpadanta (10th century), *Sanatkumāra-cariu* of Haribhadra (12th century).

Middle Indo-Aryan is characterized generally by the reduction of the complexities seen in Old Indo-Aryan. The vowel system was reduced by the merger of *ṛ* (and *ḷ*) sounds with vowels and the change of the diphthongs *ai* and *au* to the vowel sounds *e* and *o*; *e.g.*, Pāli *accha-* "bear" (Sanskrit *ṛkṣa-*), *iṇa-* "debt" (Sanskrit *ṛṇa-*), *uju-* "straight" (Sanskrit *ṛju-*), *pucchati* "asks" (Sanskrit *pṛcchati*), *mettī-* "friendship" (Sanskrit *maitrī-*), *orasa-* "breast-born, legitimate" (Sanskrit *aurasa-*). Moreover, *-aya-* and *-ava-* commonly contracted to *-e-* and *-o-*; *e.g.*, Pāli *jeti* "conquers" (Sanskrit *jayati*), *odhi-* "limit" (Sanskrit *avadhi-*). Final consonants were deleted, with the exception of *-m*, which developed to an *-ṃ* sound before which a vowel was shortened (Pāli *bhāriyaṃ* "wife"; Sanskrit *bhāryām*). Together with the trend toward replacing variable consonant stems by unchanging stems in *-a-*, this change had serious consequences for the grammar. Consonant stems steadily disappeared and were transformed to stems ending in a vowel; *e.g.*, to Sanskrit *śarad-* "autumn," *sarit-* "stream," and *sarpis-* "butter" correspond the Pāli forms *sarada-*, *saritā*, and *sappi-*. Consonant clusters were also modified in Middle Indo-Aryan; *e.g.*, Pāli *khetta-* "field" (corresponding to Sanskrit *kṣetra-*), Pāli *dakkhiṇa-* "right, south" (Sanskrit *dakṣiṇa*), *aggi-* "fire" (Sanskrit *agni-*), *puṇṇa-* "full" (Sanskrit *pūrṇa*), and *taṇhā-* "thrist" (Sanskrit *tṛṣṇā-*). The shortening of vowels before modified consonant clusters led to the use of short *ĕ* and *ŏ* sounds, which were unknown in Old Indo-Aryan; *e.g.*, Pāli *sĕmha-* "phlegm" (Sanskrit *śleṣman*), *ŏṭṭha-* "lip" (Sanskrit *oṣṭha-*).

The above phenomena are not restricted to Pāli; they are pan-Middle Indo-Aryan. Differences between Pāli and Aśokan and other Prākrits include the retention of voiceless stops (*i.e.*, *p*, *t*, *k*) between vowels in Pāli and Aśokan dialects; other Middle Indo-Aryan dialects modify them. The extreme development appears in literary Māhārāṣṭrī, in which unaspirated stops (pronounced without an accompanying audible release, or pull of breath) other than retroflexes (*ṭ*, *ḍ*) and labials (*p*, *b*) were deleted, aspirated stops (pronounced with an audible puff of breath) were replaced by *h*, retroflexes (pronounced by curling the tongue upward toward the hard palate) became voiced, and labials were replaced by *v*; *e.g.*, *loa-* "world" (Sanskrit *loka-*), *loaṇa-* "eye" (Sanskrit *locana-*), *sāhā-* "branch" (Sanskrit *śākhā-*), *paḍhai* "recites, reads" (Sanskrit *paṭhati*), and *savaha-* "curse" (Sanskrit *śapatha-*).

Characteristics of literary Māhārāṣṭrī

Essentially on the same level are the dialects of Jaina texts, but in these a *y* glide prescribed by grammarians occurs when a consonant is elided: *vayaṇa-* "face" (Sanskrit *vadana-*); *sayala-* "whole" (Sanskrit *sakala-*). In Śaurasenī, on the other hand, voiceless stops (*e.g.*, *p*, *t*, *k*) between vowels are voiced (*e.g.*, become *b*, *d*, *g*, respectively); *e.g.*, *ido* "hence" (Sanskrit *itaḥ*); *tadhā* "thus" (Sanskrit *tathā*). Though Pāli and Aśokan are at an earlier level of development with respect to these changes, they share with the rest of the Middle Indo-Aryan dialects the replacement of voiced aspirated sounds between vowels by *h*: *lahu-* "light, unimportant" from *laghu-*; *dahati* "gives" (Sanskrit *dadhāti*). Similarly, they share the change of *dy-* to *j*: *joti-* "light, brilliance" (Pāli *jotati* "shines," Sanskrit *dyotate*). Pāli and Aśokan, however, retain a *y* sound, changed to *j* in most other Prākrits; *e.g.*, the pronoun *ya-* (feminine *yā-*), as in Sanskrit, opposed to *ja-*.

The deletion of stop consonants noted above resulted in vowel sequences within words that were unknown to Old Indo-Aryan. Similarly, the extent of *sandhi* modification was restricted in Middle Indo-Aryan. The Middle Indo-Aryan vowels *ī* and *ū* do not change to *y* and *v* before dissimilar vowels in compounds; *e.g.*, Māhārāṣṭrī *rattīandhaa-* "dark of night" (Sanskrit *rātry-andhaka-*). In addition, the first of two contiguous vowels in different words is subject to deletion; *e.g.*, Pāli *manas'icchasi* (from *manasā icchasi*) "you wish in your mind."

In its grammatical system, Middle Indo-Aryan also reduced complexities. The dual number no longer exists as a separate category; for Sanskrit *dvābhyām* "by two," Middle Indo-Aryan has *dohi*(*ṃ*), with the ending *-hi*(*ṃ*) equivalent to the instrumental plural *-bhis* of Old Indo-Aryan. Among other changes is the replacement of the dative case by the genitive except in particular usages; *e.g.*, the use of forms corresponding to the Old Indo-Aryan dative to denote a purpose.

In Middle Indo-Aryan, nominal and pronominal forms are no longer strictly segregated; *e.g.*, Aśokan *vijitamhi* "in the kingdom" (also *vijite*) has a pronominal ending equivalent to Sanskrit *-smin*.

Middle Indo-Aryan verb system

In the verb system, the contrast between active (*-ti*) and mediopassive (*-te*) endings was obliterated. Further, the Old Indo-Aryan distinction between aorist, imperfect, and perfect forms was eliminated. With few exceptions, the sigmatic aorist (an aorist form with *s*) provides the only productive preterite of early Middle Indo-Aryan: Aśokan *ni-kkhamisu* "they set out" (Sanskrit *nir-a-kramiṣur*). In later Prākrits verbally inflected preterites were generally eliminated; in their place was used the past participle. For example, in Śaurasenī *devi uva-visa, mahārāo vi ā-ado* "Sit down, my queen, the king also has arrived," the past participle *ā-ado* (Sanskrit *ā-gataḥ*) agrees with *mahā-rāo* "king" (Sanskrit *mahā-rājaḥ*) in number and gender. If the verb is transitive, the participle agrees with the direct object, and the agent is denoted by an instrumental form: in Jaina Māhārāṣṭrī, *teṇa vi savvaṃ siṭṭhaṃ* "He has told everything," *teṇa* "by him" denotes the agent, and *siṭṭhaṃ* "told" (Sanskrit *śiṣṭam*) agrees with the neuter singular form *savvaṃ* (Sanskrit *sarvam*). When no object is denoted, the verb is in the neuter singular. Old Indo-Aryan used both the participial construction and the finite verb; thus to Prākrit *so vi teṇa samaṃ gao* "He also went with him" could correspond Sanskrit *so'pi tena saha gataḥ* or *so'pi tena sahāgamat* (*saha aga-mat*). The Middle Indo-Aryan development eliminated the latter.

Alternations of the Sanskrit type *as-mi*, *s-mas* were eliminated in Middle Indo-Aryan; the predominant type of present tense was formed from an unchanging vowel stem (Pāli *e-ti*, *e-nti* "go[es]").

Nominal forms of the verb system are of the same types as Old Indo-Aryan; *e.g.*, the Pāli future passive participle *kātabba-* (Sanskrit *kartavya-*) "to be done," Śaurasenī *karaṇia*; Ardhamāgadhī, Jaina Māhārāṣṭrī, and Māhārāṣṭrī *karaṇijja-* "to be done." The infinitive is commonly formed on the present tense stem, not on the root, as in Old Indo-Aryan. Thus Pāli *pappotum* is formed on the present *pappoti*; Sanskrit *prāptum* is formed on the root *prāp*, present tense *prāpnoti*.

Middle Indo-Aryan shows evidence of dialectal differentiation. The earliest documents that allow one to determine roughly the dialect distribution are Aśoka's inscriptions. These represent three major dialect areas: east, as in the inscriptions of Jaugaḍa, Dhauli, and Kālsī; west, in Girnār; and northwest, in Mānsehrā and Shāhbāzgaṛhī. Characteristic of the east dialect area is final *-e*, corresponding to *-o* in the west and *-as* in Sanskrit; in the east dialect area *l* also regularly corresponds to *r* of the west and of Sanskrit. Moreover, in the east dialect area there is a tendency to insert a vowel within consonant clusters, while in the west and northwest one of the consonants is assimilated to the other without an intervening vowel. For example, to Sanskrit *rājñas* "of the king" corresponds Girnār *rañño*, Shāhbāzgaṛhī *raño*, Jaugaḍa *lājine*. Northwest stands apart in retaining three spirant sounds, *ś*, *ṣ*, *s*, which merge to *s* elsewhere. Aśoka's eastern dialect, from the Magadha country, shows an *s* sound for Old Indo-Aryan *ś*, *ṣ*, *s*, rather than the *ś* sound typical of literary Māgadhī. Grammatical features also show dialectal variation; *e.g.*, the Aśokan dative singular form is *-āya* in the western dialects (Girnār *atthāya* "for the purpose of") but *-āye* in the east (Kālsī, Dhauli *aṭṭhāye*).

Apabhraṃśa

As noted above, the most advanced development of Middle Indo-Aryan is seen in Apabhraṃśa. Sound changes that are typical of Apabhraṃśa include the replacement of the vowel sound *a* by *u* in final syllables; *e.g.*, *karahu* "you do, make," corresponding to *karaha* (*karadha*) in other Prākrits. From stems in *-aya-* develop forms in *-au* and nasalized *-aũ* (nasalization is here indicated by a tilde): *bhaḍārau* "honored one, king" (Prākrit *bhaṭṭār-ayo*), *haũ* "I" (Aśokan *hakaṃ*). Nasalization also appears in environments in which earlier *m* occurred between vowels; *e.g.*, *gāũ* "village" (from *gāma*, Sanskrit *grāma*). Numerous other sound changes are evident, among them the development of *-s*(*s*)- between vowels into *h*: *tahŏ* "of him" (from Prākrit *tassa*, Sanskrit *tasya*); *hohinti* "will be" (compare Pāli *hossati*). Apabhraṃśa contractions, such as *-aya-* changing to *-a* and *-iya* to *-ī*, foreshadow New Indo-Aryan, in which the development was extended; *e.g.*, Apabhraṃśa *pāṇiu* "water" (Old Indo-Aryan *pāniyam*), Gujarati *pāṇī*, Hindi *pānī*.

In other points Apabhraṃśa also presaged New Indo-Aryan. The interest of Apabhraṃśa lies in the fact that contracted forms presage the New Indo-Aryan opposition of masculine, neuter, and feminine nouns; thus, Apabhraṃśa *-au*, *-aũ*, *-ī*, Gujarati *-o*, *-ũ*, *-ī* (*gayo*, *gayũ*, *gaī* "went"), Hindi *-ā*, *-ī* (*gayā*, *gaī*). The case system of Apabhraṃśa is also at a more advanced level of disintegration than that of earlier Middle Indo-Aryan, with the instrumental and locative plurals being identical in form (*-ahĩ* or *-ehĩ* for *-a*-stems) and instrumental singular forms also being used as locatives.

In the Apabhraṃśa verb system, present tense stems in *-a* predominate. Apabhraṃśa verb endings differ from those of other Prākrits. Most interesting is the 3rd person plural type *kara-hĩ* "they do," which coexists with *karanti*. The form *kara-hĩ*, corresponding to the 3rd person singular *kara-i* "he does," is formed on the model of the pair *kara-ũ* (1st person singular, "I do") and *kara-hũ* (1st person plural, "we do"). Here again Apabhraṃśa comes close to New Indo-Aryan. Moreover, Apabhraṃśa has some causative formations that do not occur elsewhere in Middle Indo-Aryan but are known from New Indo-Aryan—*bham-āḍa-i* "causes to turn," Gujarati *bhamāṛe che* "causes to turn round," and *pais-āra-i* "causes to enter," Gujarati *pesāre che* "causes to enter, to penetrate."

Apabhraṃśa syntactic patterns

Also noteworthy are two syntactic usages that closely parallel those present in New Indo-Aryan. The present participle is used as a conditional; *e.g., jai haũ mi teṇa sahũ tau karantu to kiṃ asamāhie sahũ marantu* "Even if I had performed (*karantu*) ascetic acts with him, would I have died without mental concentration?" in which the participles *karantu* and *marantu* have the value of conditionals. In Sanskrit the conditionals *a-kar-iṣya-m* and *a-mar-iṣya-m* are used; but in speaking Gujarati a person would say *jo hũ . . . karat . . . to marat,* and Hindi would have the forms *kartā . . . martā.* The Apabhraṃśa gerundive in *-iv(v)a* or *-ev(v)a* can be used as an infinitive; *e.g., pi-evae laggā* "began to drink." This is the Gujarati construction *pi-vā lāgyo* "began to drink," in which *pi-vā* is an inflected form of *pi-vũ,* that is, a verbal noun (infinitive) corresponding etymologically to the Apabhraṃśa gerundive.

Influences on Middle Indo-Aryan. In the mid-2nd century BC, the grammarian Patañjali explained that to speak faultlessly the language now called Sanskrit (as described by Pāṇini) one should imitate the correct speakers (called *śiṣṭa* "learned, educated") of Āryāvarta ("Country of the Aryans"). Earlier, the grammarian Kātyāyana (*c.* 3rd–4th century BC) had noted that Pāṇini gave lists of verb roots in order that certain Middle Indo-Aryan forms not be accepted as having been correctly derived from a Sanskrit verb root. Moreover, Patañjali noted that one should study grammar in order to learn not to use incorrect words such as *helayaḥ* instead of *herayaḥ* (a phrase used in calling to people) or *gāvī* instead of *gauḥ* "cow"; *gāvī* is a Middle Indo-Aryan word. The observations of these grammarians are considered to lend support to the view that by the 6th or 5th century BC Sanskrit as a medium of learned conversation coexisted with Middle Indo-Aryan. Further, the Pāli canon records that the Buddha enjoined his followers to use the vernaculars in communicating his teachings, and the Jaina canon identifies Ardhamāgadhī as the language to be employed for communicating the teachings of Mahāvīra. Similarly, Aśoka used Middle Indo-Aryan, not Sanskrit, in the inscriptions he ordered written throughout his kingdom; Sanskrit does not appear on inscriptions until the early centuries AD (*e.g.,* Rudravarman's inscription at Junagarh, *c.* AD 150). The coexistence of Old Indo-Aryan and Middle Indo-Aryan is to be accepted even for the time when the earliest Old Indo-Aryan texts were put to writing.

Middle Indo-Aryan shows similar evidence of the influence of linguistically more advanced vernaculars on literary compositions. The Prākrits of elegant literary compositions must have been artificial, different in many respects from the vernaculars current at the time, though reflecting languages that were current at some former time. The Old Indo-Aryan and Middle Indo-Aryan stages, then, present a picture of concurrent vernaculars with dialects and literary languages influenced by the vernaculars; it is impossible to compartmentalize the different stages as beginning and ending at any definite date.

Sources of borrowing into Indo-Aryan

The literary languages borrowed words and suffixes from earlier languages. There are Prākritisms (*i.e.,* forms of earlier Prākrits) in Apabhraṃśa; *e.g.,* the genitive singular ending *-ssa* instead of *-hõ* and 2nd person plural verb forms terminating in *-ha* instead of *-hu.* All the literary Prākrits had recourse to Sanskrit as a source for borrowing words. Words that were incorporated into the Prākrits from Sanskrit with no change in form are called *saṃskṛta-sama* "identical with Sanskrit" (or *tat-sama* "identical with that") and are contrasted with words termed *saṃskṛta-bhava (tad-bhava)* "whose origin is in Sanskrit"—that is, words that the grammarians can derive from Sanskrit by using certain rules. Another class of words, called *deśya* (or *deśī*) "belonging to the area, country," includes items that the grammarians cannot derive easily from Sanskrit and that are supposed to have been in use in particular areas from early times.

Many or most of the *deśya* words are indeed derivable from Sanskrit, but some are of Dravidian origin; *e.g., akka* "sister" (Telugu *akka*), *attā* "father's sister" (Telugu *atta*), *appa* "father" (Telugu *appa*), *ūra* "village" (Telugu *ūru*), *pulli* "tiger" (Telugu *puli*). Borrowing from Dravidian occurred also at earlier times; the Dravidians originally occupied territory much farther north than they did in Middle Indo-Aryan times. The Ṛgveda has such words as *kuṇḍa* "pitcher, pot," which is doubtless of Dravidian origin (Tamil *kuṭam* "pot"). Such borrowings become more numerous in later Sanskrit. It is not always certain that borrowing proceeded from Dravidian to Indo-Aryan, however, because Dravidian languages freely borrowed from Indo-Aryan. Thus, some scholars claim that Sanskrit *kaṭu* "sharp, pungent" is from Dravidian, but others claim that it is a Middle Indo-Aryan form deriving from an earlier **kṛt-u* "cutting" (root *kṛt*). (An asterisk [*] preceding a form indicates that it is not attested but has been reconstructed as a hypothetical form.) Whatever the judgment on any individual word, it is clear that Indo-Aryan did borrow from Dravidian, and this phenomenon is important in considering a group of sounds that sets Indo-Aryan apart from the rest of Indo-European—the retroflexes. Without doubt the influence of Dravidian is to be considered as contributing to the extension of these sounds beyond their limited occurrence in inherited Indo-European items such as *nīḍa* "nest" (from **ni-sd-o*), *iṣ-ṭa* "desired" (from **is-to*), and *stīr-ṇa* "spread out" (from **stṝ-no*). The Munda languages (or, more generally, the Austro-Asiatic languages) are also a source of some borrowing into Indo-Aryan; *e.g.,* Sanskrit *jambāla* "mud" (Santali *jobo*).

In the 8th century AD, the philosopher Kumārila mentioned not only Dravidian but also Persian and Greek as sources of foreign words. Such borrowing can be traced back to early times. In the 6th century BC Darius counted Gandhāra as a province of his kingdom, and Alexander the Great penetrated into northern India in the 4th century BC. From Iranian come words such as that meaning "inscription, writing, script"; in the northwest inscriptions of Aśoka the word is *dipi* (Old Persian *dipī*) and Sanskrit has *lipi,* the form in other Aśokan versions and in Pāli. Also from Persian is Sanskrit *kṣatrapa* "satrap"—Old Persian *xšassa-pāvan-*. Of Greek origin are such mathematical and astronomical terms as Sanskrit *kendra* "centre" (Greek *kéntron*), *jāmitra* "diameter" (*diámetron*), and *horā* "hour" (*hora*). *Yavana* "foreigner," originally the Greek word for Ionian, is known from as early as the time of Pāṇini. Later, Arabic words such as *taślī* "trigon" came into Sanskrit.

Division of India into linguistic states

The modern Indo-Aryan stage. The division of the Indian subcontinent into linguistic states and even into countries (Pakistan, Bangladesh, and India) is a recent phenomenon. Even after independence from Britain was achieved and partition had taken place, Bombay state existed until it was split into Gujarāt and Mahārāshtra states in 1960. The division of Punjab into Punjab and Haryana states in 1966 occurred as a result of Punjabi agitation for a separate linguistic state. Before independence, under British rule (entrenched from the 18th century), there were princely states within dialect areas; under Mughal rule (16th–18th centuries), Persian was the language which was used by the court and by courts of justice and this practice continued in the latter function for a time under the British. Though Hindi–Urdu may have been a lingua franca, however, the great dialectal diversity of earlier times continued.

Some of the modern Indo-Aryan languages have literary traditions reaching back centuries, with enough textual continuity to distinguish Old, Middle, and Modern Bengali, Gujarati, and so on. Bengali can trace its literature back to Old Bengali *caryā-padas,* late Buddhist verses thought to date from the 10th century; Gujarati literature dates from the 12th century (Śālibhadra's *Bharateśvara-bāhubali-rāsa*) and to a period when the area of western Rājasthān and Gujarāt are believed to have had a literary language in common, called Old Western Rajasthani. Jñāneśvara's commentary on the *Bhagavadgītā* in Old Marathi dates from the 13th century and early Maithili from the 14th century (Jyotiśvara's *Varṇa-ratnākara*), while Assamese literary work dates from the 14th and 15th centuries (Mādhava Kandalī's translation of the *Rāmāyaṇa*, Śaṅkaradeva's Vaiṣṇaviṭe works). Also of the 14th century are the Kashmiri poems of Lallā (*Lallāvākyāni*), and Nepali works have also been assigned to this epoch. The work of Jagannāth Dās in Old Oriya dates from the 15th century.

Amīr Khosrow used the term *hindvī* in the 13th century, and he composed couplets that contained Hindi. In early times, however, other dialects were predominant in the midlands (Madhyadeśa) as literary media, especially Braj Bhasa (*e.g.*, Sūrdās' *Sūrsāgar*, 16th century) and Awadhi (*Rāmcaritmānas* of Tulsīdās, 16th century). In the south, in Golconda (Andhra, near Hyderābād), Urdu poetry was seriously cultivated in the 17th century, and Urdu poets later came north to Delhi and Lucknow. Punjabi was used in Sikh works as early as the 16th century, and Sindhi was used in Ṣūfī (Islāmic) poetry of the 17th–19th centuries. In addition, there is evidence in late Middle Indo-Aryan works for the use of early New Indo-Aryan; *e.g.*, provincial words and verses are cited.

The creation of linguistic states has reinforced the use of certain standard dialects for communication within a state in official transactions, teaching, and on the radio. In addition, attempts are being made to evolve standardized technical vocabularies in these languages. Dialectal diversity has not ceased, however, resulting in much bilingualism; for example, a native speaker of Braj Bhasa uses Hindi for communicating in large cities such as Delhi.

Attempt to establish one national language

Moreover, the attempt to establish a single national language other than English continues. This search has its origin in national and Hindu movements of the 19th century down to the time of Mahatma Gandhi, who promoted the use of a simplified Hindi–Urdu, called Hindustani. The constitution of India in 1947 stressed the use of Hindi, providing for it to be the official national language after a period of 15 years during which English would continue in use. When the time came, however, Hindi could not be declared the sole national language; English remains a co-official language. Though Hindi can claim to be the lingua franca of a large population in North India, other languages such as Bengali have long and great literary traditions—including the work of Nobel Prize winner Rabindranath Tagore—and equal status as intellectual languages, so that resistance to the imposition of Hindi exists. This resistance is even stronger in Dravidian-speaking southern India. The use of English as an official language entails problems, however, because with the use of state languages for education, the level of English competence is declining. Another danger faced is the agitation for more separate linguistic states, threatening India with linguistic fragmentation hearkening back to earlier days.

Characteristics of the modern Indo-Aryan languages. The trends noted in Middle Indo-Aryan continue in New Indo-Aryan. The Middle Indo-Aryan vowel sequences *ai* and *au* were changed to single vowels during the development of New Indo-Aryan, final vowels were shortened and deleted, and *ḍ* and *ḍh* sounds between vowels were replaced by the sounds *ṛ* and *ṛh*. The noun cases were further reduced, and the introduction of nominal (noun) forms into the verb system became more pronounced.

Literary languages tend to become somewhat removed from the usual standard colloquial. Literary, or High, Hindi, for example, tends to replace some of the Perso-Arabic vocabulary with Sanskritic items, whereas literary Urdu makes great use of Perso-Arabic words. The gap is formalized in Bengali, in which a distinction is made between the highly Sanskritic language Sadhu-Bhaṣa and the colloquial standard called Calit-Bhasa.

Phonology. [Note: The forms of the words given below reflect actual pronunciation, rather than being transliterated versions of the standard orthographies. For New Indo-Aryan the symbols *ə*, pronounced as the *a* in English "sofa," and *a* are used for the sounds earlier transcribed as *a* and *ā*, respectively; *e.g.*, Gujarati *karũ* "I do" and *māro* "beat" are now written *kərũ* and *maro*. This practice permits certain contrasts to be made among sounds that are significant in the description of dialectal features. In Kashmiri words, *a* is short, opposed to *ā*.]

Vowel changes in New Indo-Aryan

Vowels in sequence contracted in early New Indo-Aryan; *e.g.*, Old Indo-Aryan *aśīti* became Middle Indo-Aryan *asīi*, Hindi and Punjabi *əssī*, and Bengali *aši* "80." Further, *ai* and *au* sounds changed to *e* and *o*, and *aũ* to *ũ*, while *iu developed into ī*. The diphthongs *ai* and *au* were retained well into the New Indo-Aryan period and are still pronounced in some areas; *e.g.*, Braj Bhasa *kərəũ* "I do," *kərəi* "he does." Middle Indo-Aryan *-ḍ-* and *-ḍh* developed into the flaps *ṛ* and *ṛh*; *e.g.*, Prākrit *sāḍiā* "woman's garment," Kashmiri, Lahnda, Hindi, Gujarati, Bhojpuri, Bengali, Oriya *saṛī* "sari"; and Prākrit *paḍh-* "recite, read," Sindhi *pəṛh-əṇu*, Lahnda *pəṛh-əṇ*, Hindi, Punjabi *pəṛh-na*, Gujarati *pəṛh-vũ*, Marathi *pəṛh-ṇə* "study."

Stress is not generally contrastive in New Indo-Aryan as it is, for example, in English (*e.g.*, noun "éxport," verb "expórt"), though different areas have different rules for placing major emphasis on a given syllable. For example, in Hindi, in which vowel length is pertinent, *gilá* "swallowed" has major stress on the last syllable, *gī́la* "wet," on the first. In Gujarati, on the other hand, vowel length is not pertinent; the stress position depends on which vowels occur in contiguous syllables and on the structure of the syllables, whether open or closed; *e.g.*, *júno* "old," but *dukán* "store." In Bengali each syllable of a word receives about equal stress.

The sounds that most clearly distinguish Indo-Aryan from the rest of Indo-European are the voiced aspirate stops (*gh* and the like, pronounced with an accompanying audible puff of breath) and the retroflexes (*ṭ* and so on, pronounced by curling the tongue upward toward the hard palate). In the outlying New Indo-Aryan areas, however, the sound system is reduced. Sinhalese has no aspirated stops, Assamese has no retroflexes, and Kashmiri has no voiced aspirates. The geographic position of these languages doubtless contributed to these losses: Sinhalese coexists with Tamil, Assamese is surrounded by Tibeto-Burman languages, and Kashmiri is on the border of the Iranian area.

Early New Indo-Aryan dialect distribution

New Indo-Aryan shows evidence of early dialect distribution; this is discernible by considering sound changes proper to each group. The eastern group (Assamese, Bengali, Oriya) has three important changes. Long and short *i* and *u* merged; *e.g.*, Assamese *nila*, Oriya *niḷɔ* (ɔ is similar to the *o* of "coffee" in some English dialects), Bengali *nil* "blue-black" but Sanskrit *nīla*; Assamese *dhuli*, Bengali *dhulo*, Oriya *dhuḷi* "dust" but Hindi *dhūl* and Sanskrit *dhūli*. The vowel sound *a* of Middle Indo-Aryan was replaced by ɔ in Bengali and Oriya and ɒ (similar to the *o* of "hot" in southern British English) in Assamese in initial position and open syllables; *e.g.*, Bengali *mɔron*, Oriya *mɔrɔn*, Assamese *mɒrɒn* "death"; Sindhi, *mərəno* "mortal, death," Sinhalese *mərəṇə*, Gujarati, Marathi *mərəṇ* (compare Sanskrit *maraṇa-*). Moreover, in this group a vowel is affected by the quality of the vowel in a following syllable. For example, in Bengali *ami kori* "I do," the verb root has *o* followed by *i* in the next syllable, but *tumi kɔro* "you do" has an ɔ sound; similarly, *ami kini* "I buy" but *tumi keno*. As a result of vowel assimilation also, Assamese has an ɔ sound instead of ɒ representing Middle Indo-Aryan *a*: Assamese *xɔhur*, Bengali *šošur* "husband's father" (compare Hindi *səsur*, Prākrit *sasura-*, Sanskrit *śvaśura-*).

Assamese and Bengali are set off from Oriya. In the former two, Middle Indo-Aryan *ḍ* and *ḍh* merge medially to *ḍ* (then *ṛ*) with a subsequent development to *r* in Assamese; *e.g.*, Oriya *daṛhi*, Bengali *daṛi*, Assamese *dari* "beard"; Hindi, Gujarati *daṛhī*, Prākrit *dāḍhiā*. Assamese is also distinguished from Bengali by several developments, among them the merger of Assamese retroflex sounds with dental sounds; *e.g.*, Assamese *ut* "camel" but Bengali *uṭ*, Oriya *oṭɔ*, Sindhi *uṭhu*, Lahnda, Pahari *uṭṭh*, and so on. Assamese also has *s* for earlier *c* and *ch* sounds and a *z* sound for *j* and *jh;* *e.g.,* Assamese *kas* "glass," Bengali *kac;* Assamese *azi* "today," Oriya *aji,* Bengali, Hindi *aj.* In addition, Assamese replaced an *s* sound initially by *x* and between vowels by *h*—*xɔhur.*

Particular sound changes also characterize languages of the northwest. In this group, an older voiceless stop (*e.g.,* *-t*) became voiced (*e.g.,* became *d*) after a nasal sound; in other areas, the voiceless stop is retained: Kashmiri *dand,* Punjabi *dənd,* Sindhi *ḏəndu* "tooth" (the *ḏ* in Sindhi is an imploded stop; see below) but Assamese, Bengali, Hindi, Gujarati, Marathi *dā̃t,* Sinhalese *dətə* (Sanskrit *danta-*). Moreover, in the northwest group a voiced stop (*e.g.,* *d*) preceded by a nasal was assimilated to the latter, resulting in two nasals, which were subsequently reduced to one in some areas; in the rest of New Indo-Aryan, the vowel preceding the nasal was nasalized. Thus, Kashmiri *don* "churning stick," Sindhi *ḏənu* "tribute," Punjabi *dənn* "fine," Lahnda *ḏənn* "force," Kumauni *dan* "roof" contrast with Assamese *dā̃r* "pole," Bengali *dā̃ṛ* "oar," Hindi *dā̃ḍ* "oppression, fine," and others; all forms derive from Old Indo-Aryan *daṇḍa-* "stick, staff, club, royal power, fine, punishment."

In the sequence of a short vowel followed by two consonants, Pahari differs from the rest of the northwest group and agrees with the rest of New Indo-Aryan. In the northwest this sequence either remained unchanged or the cluster was simplified without lengthening of the vowel; other languages generally simplified the cluster and lengthened the vowel: Punjabi *bhətt,* Sindhi *bhətu,* Lahnda *bhət,* Kashmiri *batɨ* "cooked rice, food" but Nepali, Kumauni, Hindi, Assamese, Bengali, Gujarati, Marathi *bhat.*

Sound changes in Dardic

Dardic occupies a special position. The sibilant sounds did not all merge here. For example, Kashmiri, a Dardic tongue, has *šurah* "16" with *š* rather than *s,* as in most other Indo-Aryan languages, and *sat* "7" with *s.* Further, voiced aspirated stops merged with unaspirated stops in Dardic; *e.g.,* Kashmiri *gur* "horse" but Hindi *ghoṛa;* Kashmiri *dɔd* "milk" but Hindi *dūdh.*

One major feature distinguishing Sindhi from the rest of the northwest group is the development of a series of imploded stops (also called suction stops and recursive stops), for *b, ḏ, j,* and *g.* Implosive stops also occur in the Sindhi vicinity; for example, Kacchi has imploded *b.* Another feature that distinguishes Sindhi from other northwest languages, including Kacchi, is the retention of the Middle Indo-Aryan final short vowels; *e.g.,* Sindhi *əkhi* "eye" but Hindi *ā̃kh* (Middle Indo-Aryan *akkhi-*).

Punjabi is distinguished from other members of the northwest group by its tonal system, having low (̀), mid (̄), and high (́) tones. Initial voiced aspirated stops of earlier Indo-Aryan appear in Punjabi as voiceless stops with low tone on the following vowel; *e.g.,* Punjabi *kòṛa* but Hindi *ghoṛa;* Punjabi *tàī* "2½" but Hindi *ḍhaī.* Non-initially, a voiced aspirate became unaspirated and the preceding vowel received high tone; thus, Punjabi *dúd* "milk" but Hindi *dūdh,* and Punjabi *láb* "profit" but Hindi *labh.*

Gujarati, Marathi, and Konkani in the west and southwest differ from the languages of the midlands in that, as in the east, there is no contrast between long and short *i* and *u* vowels. The *i* of Gujarati and Marathi *vis* "20" is pronounced like the *ee* of English "teeth," the *i* of Gujarati *iccha* and Marathi *iččha* "wish" like the *i* of "pitch," but such a difference is not contrastive, as it is in Hindi (*gīla* "wet": *gila* "swallowed"). Gujarati has certain features that, in turn, set it apart from the other languages of this group. In addition to *e* and *o* sounds, it has the open vowels ɛ, ɔ; *e.g.,* *cɔthũ* "fourth" (Middle Indo-Aryan *cauttha*), *bɛs-vũ* "to sit" (Middle Indo-Aryan *baisai* "sits"). Moreover, Gujarati has murmured vowels, generally developed from vowels followed by *h;* *e.g.,* *kɛh che* "says" (*h* represents murmuring of the vowel), Old Gujarati *kahai chai.* Marathi and Konkani have two series of affricate sounds; *e.g.,* *č* (pronounced as the *ch* in English "chat"; the equivalent of *c* in some other languages) and *c* (pronounced as the *ts* of "rats").

There was clearly mutual influence of Indo-Aryan languages at an early time, together with movement of groups of speakers (compare the position of Pahari). Thus, while Punjabi *səcc* "true" is the expected form comparable to Middle Indo-Aryan *sacca-* (Old Indo-Aryan *satya-*), Hindi *səc* "true" does not represent the expected outcome. The item *səc* must come from the Punjabi area.

Gender in New Indo-Aryan

Grammar. Like Middle Indo-Aryan, New Indo-Aryan distinguishes only two numbers—singular and plural. Unlike Middle Indo-Aryan, the New Indo-Aryan languages differ in the degree to which gender distinctions are made. Three genders are retained in the west and southwest (Gujarati, Marathi, Konkani), and this is true also of Sinhalese. Unlike Gujarati, Marathi, and Konkani, in which every noun, whether it denotes an animate being or not, has a particular gender that is unpredictable, Sinhalese restricts masculine and feminine gender to animates and neuter to inanimates. The eastern group (Assamese, Bengali, Oriya) has no grammatical gender distinctions, and two genders are distinguished elsewhere.

Over a large area of New Indo-Aryan the noun has only two cases—direct and oblique. A lack of distinction between direct and oblique cases in the plural is typical of several languages, including forms in Hindi, Gujarati, Marathi, and Bhojpuri. Direct forms are used independently, oblique forms before postpositions (words or word elements following a noun that function similarly to English prepositions) and other affixes; the combination of stem and postposition serves the function of inflected case forms of earlier Indo-Aryan. Thus, to denote an object (direct or indirect) Hindi uses the postposition *ko,* which occurs in direct object constructions normally only with nouns denoting animate beings; *e.g.,* *ləṛke-ko dekh-ta hɛ* "He sees the boy," *ləṛke-ko miṭhaī do* "Give a sweet to the boy." Other postpositions are *mẽ* "in," *pər* "on," *se* "from, with, by means of." A large group of postpositions are linked to the noun with the affix *ka* (oblique form *ke,* feminine *kī*), which also is used to form adjectives (possessives); *e.g.,* *ləṛke-ke sath gəya* "He went with the boy," *ləṛke-ke pas hɛ* "The boy has it" (literally, "It is by the boy"). Many such postpositions represent old nominal (noun) forms. Other New Indo-Aryan languages have systems similar to that of Hindi, though the forms of the postpositions differ.

Though the nominal (noun) system of Punjabi is very close to that of Hindi, it has separate ablative (indicating separation and source) and locative (indicating place) forms in the singular and plural, respectively, for nouns such as *koṭha* "house"; *e.g.,* *koṭhiõ* "from the house," *koṭhī̃* "in the houses." Some languages have a fuller case system than that noted above; *e.g.,* Bengali has a genitive singular ending, a genitive plural ending, and a locative case. Similarly, Kashmiri has nominative, dative, ablative, and agentive cases. Not all such case forms are inherited from Middle Indo-Aryan. In addition to case endings, these languages also use postpositions; *e.g.,* Kashmiri *garājas-andar* "in the garage," with *-andar* after the dative ending *-as.*

Comparatives and superlatives

Adjectives behave generally in the same way as nouns but have a syntactic restriction. In Hindi the possessive is in the oblique (non-nominative) form, as is the noun after which it occurs; but in the plural, only the noun has the oblique form. Further, the formation of comparatives and superlatives with derivative affixes has been eliminated. To a Sanskrit sentence such as *ime amū-bhyaḥ āḍhya-tarāḥ* "These (people) are richer than those," in which the comparative *āḍhya-tara* occurs construed with the ablative form, corresponds a Hindi sentence *ye un-se əmīr hɛ̃,* in which no comparative affix is used—literally, "These are rich from (*i.e.,* in comparison with) those." Comparable constructions with a postposition meaning "from" occur elsewhere in New Indo-Aryan.

The pronominal system of New Indo-Aryan formally resembles the Middle Indo-Aryan stage more than its noun system. For example, Gujarati *hũ* "I," *mɛ̃* "I" (agentive), *əme* "we" (also agentive) are directly comparable to Apabhraṃśa *haũ, maĩ, amhaĩ.* The number distinctions of the Middle Indo-Aryan pronoun have been replaced, however, by distinctions of familiarity and politeness. For example, Hindi and Bengali have a three-way distinction—Hindi *ap,* Bengali *apni* "you" are polite or honorific forms; Hindi *tum,* Bengali *tumi* are informal forms; and Hindi *tũ,* Bengali *tui* are used only for inferiors and small children. (Hindi and Bengali differ, however, in the plural forms of these.) In Gujarati, on the other hand, *tũ* is a very familiar pronoun, whereas *təme* is used generally, covering the approximate domains of Hindi *ap* and *tum; ap,* if used, strikes the hearer as fawning. Marathi has a similar system. Southwestern languages also make a distinction in the 1st person plural between inclusive and exclusive, the exclusive excluding the person spoken to. In the form of the relative pronoun and the 3rd person pronoun, languages differ in the degree to which gender distinctions are made, thus contrasting with Old and Middle Indo-Aryan, in which these forms had three genders. For example, Marathi has masculine, feminine, and neuter for the relative pronoun, while Bengali has animate and inanimate.

New Indo-Aryan languages differ in the degree to which finite verb forms have been replaced by nominal (noun) forms. In Bengali a contrast is made between continuous or actual present (English "be . . . -ing") and noncontinuous or habitual present; *e.g., ami kaj kor-i* "I work" (literally, "I do work"), with the ending *-i,* contrasts with *ami kaj kor-ch-i* "I am working," in which *ch* intervenes between the root and the ending. Hindi has a similar contrast but uses nominal forms; *e.g., mɛ̃ kam kar-ta hũ* "I work," *mɛ̃ kam kər rəh-a hũ* "I am working." Both contain the finite form *hũ* of the auxiliary; but *kər-ta* and *rəh-a* are nominal forms, the latter the past of *rəh-*"stay." Gujarati has both types, the present tense using finite verb forms, the imperfect employing nominal forms; *e.g., hũ kam kərũ chũ* "I work, am working" and *hũ kam kər-to hə-to* "I was working, used to work." Even in areas in which finite forms are not used in the present, they occur in the imperative forms and what may be called the subjunctive; *e.g.,* Hindi *tum kam kər-o* "work," *mɛ̃ əndər aũ* "May I come in?"

Person–number system

The person–number system of the New Indo-Aryan verb accords with the use of pronouns. For example, the forms *ja-o, kər-o* in Gujarati *təme kyã jao cho* "Where are you going?" and *šũ kəro cho* "What are you doing?" are historically plurals but are used with reference to one person addressed by the pronoun *təme.* Similarly, in Hindi, in which a person distinction is not made in the plural, *ap kəhã ja rəhe hɛ̃, ap kya kər rəhe hɛ̃,* equivalent in meaning to the Gujarati sentences, have the plural form *rəhe hɛ̃.* Bengali has completely given up any number distinction in verb forms: *ami/amra kori* "I/we do." In the 3rd person a distinction is made between ordinary and honorific: *še* (ordinary)/*tini kɔren,* plural *tara/tãra kɔren.* Other languages (*e.g.,* Hindi) also have honorific forms, for which the plural is used.

In the formation of the future there are again regional differences. Some retain the future in *-s-* (Gujarati *hũ kər-iš,* 3rd person *e kər-š-e*) or *-h-* (*e.g.,* eastern dialects of Braj Bhasa, *cəlihɔ̃* "I will go"). Characteristic of the Eastern languages and of Bihari (including Bhojpuri, Magahi, Maithili) is the suffix *-b-; e.g.,* Bengali *jabe* "will go." All of these are finite forms. On the other hand, in Hindi and adjoining areas, the future is inflected for gender.

A similar contrast between the use of verbal and nominally inflected forms also appears in the past tense forms. The predominant pattern in New Indo-Aryan is that of Middle Indo-Aryan: forms are used that are etymologically participles.

The New Indo-Aryan languages retain the passive and causative forms. The causative is conservative in retaining both the affixes that appear in Middle Indo-Aryan and vowel alternation. The passive is also formed by affixation in some areas. But many languages also have a compound formation involving the verb *ja* "go" and an auxiliary (*hɛ̃*); *e.g.,* Hindi *yəhã hindī bol-ī ja-t-ī hɛ̃* "Hindi is spoken here."

There are other auxiliaries, which, like *hɛ̃,* can occur with any verb in the language; *e.g.,* the verb "can," Hindi *sək-,* Gujarati *šək.* A characteristic feature of New Indo-Aryan, however, is the use of certain verbs, variously called vector verbs or compound verbs, in restricted contexts and with particular semantics. For example, one can say *mər gə-ya* "He died," *bhūl gə-ya* "He forgot," *bol uṭh-a* "He blurted out" in Hindi, using the verbs *ja* "go" (masculine singular past *gə-ya*), *uṭh* "stand up." This phenomenon is pan-Indo-Aryan and still requires investigation.

New Indo-Aryan word order

The examples cited above also illustrate the normal word order in New Indo-Aryan languages: subject (including agential forms), object (with attributive adjectives preceding), verb (together with auxiliaries). Adverbials can precede the full sentence or occur after the subject, with slight differences in emphasis; *e.g.,* Hindi *mɛ̃ kəl aũga,* or *kəl mɛ̃ aũga* "I will come tomorrow (*kəl*)." Relative clauses normally precede correlatives: Hindi *jo admī kəl tumhare ghər-mẽ tha vo kɔn hɛ* "Who (*kɔn*) is the man (*admī*) who (*jo*) was in your house yesterday?" A notable exception to the normal final position for verbs occurs in Kashmiri, in which the verb usually occurs in second position after the subject; thus, to Hindi *vo kha rəha hɛ* "he is eating" corresponds Kashmiri *su chu khavān* with the auxiliary *chu* after the subject.

Vocabulary. The two most important sources of non-Indo-Aryan vocabulary in New Indo-Aryan are Persian (including Arabic items introduced through Persian), the court language of the Mughals, and English. The Perso-Arabic vocabulary permeates every aspect of New Indo-Aryan vocabulary, especially in the midlands (Uttar Pradesh through the Punjab). There are, of course, Hindi-Urdu words proper to Islām: Hindi *kuran* "Qur'ān," *'īd* (name of a holy day), *nəmaz* (certain prayers), *məsjid* "mosque," as well as the word for "religion," *məz̈həb.* In addition, there are numerous Perso-Arabic military and administrative terms (*kila* "fort," *səvar* "horseman," *ədalət* "court of justice"); architectural and geographic terms (*imarət* "building," *məkan* "house," *məhəl* "palace," *duniya* "world," *ilaka* "province"); words having to do with learning and writing (*kələm* "pen," *kitab* "book," *ədəb* "literature, good manners") and with apparel (*jeb* "pocket," *moja* "socks," *rumal* "handkerchief") and anatomy (*khūn* "blood," *gərdən* "neck," *dil* "heart," *bazu* "arm," *sər* "head"). Indeed some of the most common vocabulary is of this origin: *tārīkh* "date," *vəkt* "time," *sal* "year," *həfta* "week," *umər* "age," *admī* "man," *ɔrət* "woman," and others. Even the grammatical apparatus of postpositions and conjunctions reflects Perso-Arabic influence; *e.g., -ke bad* "after," *əgər* "if," *məgər* "but," *ya* "or."

The colloquial language used by any Hindu or Muslim communicating in Hindi-Urdu will contain a large number of such words. There have been efforts to polarize the two, and at times champions of Indo-Aryan have tried to replace Perso-Arabic vocabulary with Sanskritic words. The style that tends toward eliminating all but the most common Perso-Arabic words may be called High Hindi, written in the Devanāgarī script, as opposed to High Urdu, which retains Perso-Arabic of long standing, uses Persian and Arabic for learned vocabulary and is written in the Perso-Arabic script.

Influence of English on vocabulary

The influence of English as a source of borrowing still continues, and it is rare to hear a conversation on any technical subject among speakers of any Indian language in which English words are not liberally used. Among loanwords from English are names of conveyances such as Hindi *rel-gaṛi* "railroad-train" and *ṭɛksī* "taxi"; profession names such as *injinīr* "engineer," *jəj* "judge," *ḍaktər* "Western doctor," *pulis* "police"; and terms of educational administration such as *kaləj* "college" and *yunivərsiṭī* "university." English words are susceptible to replacement in India by Sanskritic ones as are those of Perso-Arabic origin.

Of much lesser magnitude are New Indo-Aryan borrowings from other languages, among them Portuguese and Turkic. From the latter, the word *urdū* came to be used as the name of a language. From Portuguese come such Hindi words as *ənənnas* "pineapple," *paũ* "(Western style) bread," *kəmīz* "(Western) shirt," *kəmra* "room," and *girja* "(Christian) church."

Writing systems. Ancient India had two main scripts in which Indo-Aryan languages were written. Kharoṣṭi, used in the northwest, is of Aramaic origin and is written from right to left; Brāhmī, of North Semitic origin, is written from left to right and appears earliest on Aśokan inscriptions in areas other than the northwest. Most scripts of New Indo-Aryan are developments of the Brāhmī. The Devanāgarī (or simply Nāgarī), used for writing Sanskrit documents in North India, is the script of Hindī and Marāṭhī as well as Nepālī. Gujarātī uses a more cursive derivative. Devanāgarī also is used, mainly among Hindus, for Kashmirī, which has, in addition, a traditional script called Sarada, which is not now in common use. The Perso-Arabic script is used instead. Also usually written in Perso-Arabic are Urdū and Sindhī (for which the Devanāgarī also is used in schools in India), whereas Punjabi employs it in Pakistan as well as a particular script of its own, known as Gurmukhi ("From the Teacher's Mouth") in the sacred writings of the Sikhs. In the east, the scripts used for Bengali and Assamese are closely related; and that of Oṛiyā, related to the other two, is highly cursive like that of neighbouring Dravidian languages. Such is also the case with Sinhalese.

The traditional alphabets are both over-explicit and not clear enough with regard to accurate representation of the spoken word. As systems in which a consonant symbol with no other accessory symbol accompanying it stands for the syllable consisting of the consonant followed by short *a,* they require previous knowledge of items for correct interpretation; Hindī *kərta* is written *ka-ra-tā* in the Devanāgarī, and, to pronounce it properly, one must know that the word has only two syllables. Although Bengali has only the spirant sound *ś,* the alphabet has symbols for *ś, ṣ,* and *s,* as in Old Indo-Aryan; but verb forms such as *kori* and *kəren* are written *ka-ri* and *ka-re-na,* both with the same initial symbol. And, though syllabic *ṛ* was lost as early as Middle Indo-Aryan, the scripts have a separate symbol for this. Script reform has been suggested; it has even been proposed that all Indo-Aryan languages adopt a Latin (roman) alphabet with diacritics, but chances for this are poor. (Ge.Ca.)

THE IRANIAN LANGUAGES

Languages of the group. The various Iranian languages fall distinctly into three categories—Ancient, Middle, and Modern Iranian.

Ancient (Old) Iranian. Of the ancient Iranian languages, only two are known from texts or inscriptions, Avestan and Old Persian, the oldest parts of which date from the 6th century BC. Avestan was probably spoken in northeastern Iran, and Old Persian is known to have been used in southwestern Iran. Other ancient Iranian languages must have existed, and indirect evidence is available concerning some of these. Thus, from the 5th-century-BC historian Herodotus, the Median word for "female dog" (*spaka*) is known, and a number of Median loanwords have been recognized in the Old Persian inscriptions. In addition, a number of Median personal names are attested in various sources. It is likely that all those languages that are known only from the Middle Iranian period were in fact spoken in a less developed form in the ancient period. It is possible that the same observation applies to some of those modern Iranian languages that are not attested in the earlier periods.

Avestan and Old Persian

The degree of mutual intelligibility that existed among the ancient Iranian languages is not known with certainty. The differences in the nature of the surviving sources have to be borne in mind. On the one hand, there is the religious poetry of Zoroaster in the Avestan language and, on the other, the official inscriptions of the Achaemenid rulers in Old Persian. Differences in the method of transmission present a further difficulty in the way of direct comparison. Nevertheless, it can safely be stated that the degree of mutual intelligibility must have been much greater between the ancient languages than between the Middle Iranian languages and that those languages geographically closer to each other probably were mutually understood better than those spoken in areas farther apart.

Avestan can hardly be said to be known beyond the ancient period, although only the earliest texts, the *Gāthās,* are as old as the 6th century BC, and the later texts represent the language of several subsequent centuries. Old Persian, on the other hand, itself spanning the 6th to the 4th century BC, was continued more or less directly by the various forms of Middle Persian. Even in this case, however, although both Old and Middle Persian represent the language of the royal court, the considerable differences between them remain unexplained.

Middle Iranian. Middle Persian is known in three forms, not entirely homogeneous—inscriptional Middle Persian, Pahlavi (often more precisely called Book Pahlavi), and Manichaean Middle Persian. Middle Persian belongs to the period 300 BC to AD 950 and was, like Old Persian, the language of southwestern Iran. In the northeast and northwest the language spoken was Parthian, which is known from inscriptions and from Manichaean texts. There are no significant linguistic differences in the Parthian of these two sources. Most Parthian belongs to the first three centuries AD.

Middle Persian and Parthian were doubtlessly similar enough to be mutually intelligible, but they differ so greatly from the eastern group of Middle Iranian languages that these must have appeared to be almost foreign languages. The languages of the eastern group, moreover, cannot have been themselves mutually intelligible. The main known languages of this group are Khwārezmian (Chorasmian), Sogdian, and Saka. Less well-known are Old Ossetic (Scytho-Sarmatian) and Bactrian, but from what is known it would seem likely that these languages were equally distinctive. There was probably more than one dialect of each of the languages of the eastern group, although there is certainty only in the case of Saka, for which at least two dialects are clearly attested. The main Saka dialect is known as Khotanese, but a small amount of material survives in a closely related dialect called Tumshuq, formerly known as Maralbashi.

Eastern group of Middle Iranian languages

A few words are known in all of these eastern Iranian languages from as early as the 2nd to the 4th century AD, but substantial evidence begins for Sogdian in the 4th century, for Saka probably no earlier than the 7th century (though that for Tumshuq may be a few centuries older), and for Khwārezmian not until the 12th century and later. The principal evidence for Bactrian belongs to the 2nd century. To the same period belong the Scytho-Sarmatian names of the earliest inscriptions.

All the eastern Iranian languages of the Middle Iranian period were spoken in Central Asia, with the exception of the language of the Scytho-Sarmatian inscriptions from what is now Ukraine, north of the Black Sea. More precisely, Bactrian was spoken in northern Afghanistan and in the adjacent parts of Central Asia. Khwārezmian was the language of Khwārezm, a historic region in present-day Turkmenistan and Uzbekistan but formerly of greater extent. Scholars believe that Sogdian was probably spoken over most of Central Asia, especially in eastern Uzbekistan, Tajikistan, and western Kyrgyzstan. There were also colonies of Sogdians in various cities along the trade routes to China; in fact, most Sogdian material comes from outside Sogdiana. The Saka dialects, Khotanese and Tumshuq, were spoken in Chinese Turkistan, modern Sinkiang; Tumshuq is the name of a small village in the extreme west of Sinkiang. Khotanese was spoken in Khotan near the modern city of Khotan (Chinese Ho-t'ien [Hotan]) on the southern route across the Takla Makan Desert and within about 100 miles (160 kilometres) to the north and to the east of Khotan, where manuscripts have been found, mainly at the sites of former shrines and monasteries.

Modern Iranian. The discontinuity already observed between Old and Middle Iranian is even more striking between Middle and Modern Iranian. There are no modern counterparts to Khwārezmian, Bactrian, and Saka, and there is no direct continuity in the case of any of the other Middle Iranian languages. Even Modern Persian does not represent a straightforward continuation of Middle Persian but is rather a koine (a dialect or language of a small area that becomes a common or standard language of a larger area), based mainly on Middle Persian and Parthian but including elements from other languages and dialects.

Although Sogdian is known in several forms, possibly representing different dialects, none of these can be considered the direct ancestor of modern Yaghnābī, spoken at present in the valley of the Yaghnob River, a tributary of the Zeravshan. Yaghnābī, nevertheless, certainly belongs linguistically to the Sogdian family. Similarly, the languages of the Scytho-Sarmatian inscriptions may represent dialects of a language family of which Modern Ossetic is a continuation, but it does not simply represent the same language at an earlier date.

Modern Iranian state languages

Only four of the many modern Iranian languages are the official languages of the state in which they are spoken. The chief of these is Persian (known in Persian as Fārsī), the national language of Iran, which is spoken by about 27,000,000 people as a native language. A dialect of Persian known as Dari is recognized, moreover, as a second language in Afghanistan. The national language of Afghanistan is the East Iranian language known as Pashto, of which there are some 9,000,000 speakers, many living in Pakistan. Tajik is spoken by at least 7,000,000 people widely spread throughout Tajikistan and the rest of Central Asia and is readily intelligible to speakers of Persian, to which it is very closely related, although it is in some respects more archaic.

In addition to being the national language of Tajikistan, Tajik is important as the lingua franca of the Pamirs mountain range, a region where a remarkable variety of Iranian languages and dialects are spoken. Some 700,000 people speak Ossetic. Most of the Ossetes live in North Ossetia in Russia and South Ossetia in Georgia. Although spoken in the heart of the Caucasus Mountains, Ossetic is an East Iranian language not mutually intelligible with any other Iranian language.

Two other Iranian languages, Kurdish and Balochi (Baluchi), are spoken over a vast area, although they have not been officially accepted as the national language of an established state. Kurdish is spoken by more than 10,000,000 people living in Iran, Iraq, Turkey, Syria, and Transcaucasia. More than 5,000,000 people speak Balochi as their chief language; they are spread widely over parts of eastern Iran, Pakistan, Afghanistan, and Central Asia. In Iran, Balochi speakers live mainly in Baluchistan, a region in the southeast that now forms part of a province with Sīstān. In Pakistan, Balochi speakers live mainly in the southwestern province of Balochistān; in Central Asia, they are found mainly around Mary (Merv) in southern Turkmenistan; and in Afghanistan, they are widely scattered, mainly over the southwestern portion of the country. There is a sizable Balochi colony in Oman, and many Balochi merchants have settled in the sheikhdoms of southern Arabia and along the east coast of Africa as far south as Kenya. Linguistically, Balochi and Kurdish are both West Iranian languages. Balochi is thus much more closely related to Kurdish than it is to its close neighbour Pashto. According to the most likely theory, the present eastern location of Balochi speakers is the result of migrations from the region of the Caspian Sea during the Middle Ages.

Dialects. The six modern Iranian languages discussed above are the only ones that have an established literary tradition. They are not, however, homogeneous, each having its own dialect divisions. No definitive dialect classification has yet been made, nor indeed has any attempt at systematic classification of the whole range of Iranian languages won wide acceptance. The usual practice, followed here, is simply to list the main languages in groups of varying size, arranged on a roughly geographic basis.

Ossetic dialects

There are two main dialects of Ossetic: the eastern, known as Iron, and the western, known as Digor (Digoron). Of these, Digor is the more archaic, Iron words being often a syllable shorter than their Digor counterparts—*e.g.*, Digor *madä*, Iron *mad* "mother." Iron is spoken by the majority of Ossetic speakers and is the basis of the literary language. Chosen in the 19th century for the translation of the Bible, it is still the official language today. Little is known of the other Ossetic dialects. A small amount of the Ossetic dialect of Tual in the south, which differs little from Iron, was published in Georgian script at the beginning of the 19th century.

Yaghnābī is still spoken by a small number of people southeast of Samarkand, Uzbekistan. It has two main dialects, eastern and western, which differ only slightly. The characteristic difference is between a western *t* sound and an eastern *s* sound from an older *θ* sound (as *th* in English "thin")—*e.g.*, western *mēt*, eastern *mēs* "day," beside Sogdian *mēθ* (Christian Sogdian *myθ*).

Dialects of the Shughnī group are spoken in the Pamirs. Closely related to this group is Yāzgulāmī. A period of a Yāzgulāmī-Shughnī common language (protolanguage) has been postulated by some scholars, after which it separated first into Yāzgulāmī and Common Shughnī; and then Common Shughnī gradually divided into Sarīkolī, Oroshorī-Bartangī, Roshānī-Khufī, and Bajuvī-Shughnī. Sarīkolī, the easternmost of these dialects, is spoken in northwestern China.

Speakers of Wakhī number 10,000 or so in the region of the upper Pyandzh (Panj) River. Vākhān (Wākhān), the Persian name for the region in which Wakhī is spoken, is based on the local name Wux̌, a Wakhī development of *Waxšu, the old name of the Oxus River (modern Amu Darya). (An asterisk denotes a hypothetical, unattested, reconstructed form or word.) The Wakhī language is remarkably distinct from its neighbours and has many archaic features.

Around the bend of the Amu Darya and in the valley of the Vardūj River to the southeast, a few people speak dialects of the Sanglechī-Ishkashmī group. This group is clearly distinguished from its neighbours but is closely related to the other languages of the Pamirs.

Some 6,000 people speak dialects of the Yidghā-Munjī group. Monjān is a very remote valley located in northern Afghanistan, and it is separated by a mountain pass from the Sanglechī-speaking region. Yidghā is spoken in the valley of the Lutkho River and in the nearby city of Chitrāl, a region now in Pakistan. Yidghā-Munjī is most closely related to Pashto.

Pashto dialectal groups

The existence of two dialectal groups within Pashto has long been known. Thus, the word Pashto represents a southwestern dialect form (*paštō*), in contrast to a northeastern (*paxtō*). According to one hypothesis, Pashto literature, which exists certainly from the 17th century and possibly from the 11th, was created among the northeastern tribes. Two minor dialects, Wazīrī and Waṉētsī, have some features of special interest.

Although spoken in a few villages in Afghanistan, two languages have features closely associating them with Western Iranian. These are Parāchī, spoken in the Hindu Kush north of Kabul, and Ormurī, found in two dialects, one in the Lowgar River valley south of Kabul and the other in Kāniguram in Wazīristān.

Farther south is the wholly West Iranian language Balochi, mentioned above. Despite the vast area over which Balochi is spoken, its numerous dialects are all mutually intelligible. The most recent study of the Balochi dialects divides them into six groups: Eastern Hill dialects; Rākhshānī dialects including that of Mary; Sarawānī; Kechī; Loṭunī; and the coastal dialects. Of these, Rākhshānī is the most widely spoken and is used for broadcasting both in Pakistan and in Afghanistan, but the coastal dialects have the greatest prestige and the most extensive literature.

In the southeastern corner of Iran, Balochi gradually gives way to the Bashkardī dialects.

In central Iran the influence of Modern Persian is everywhere strongly felt, and it is often difficult to distinguish between dialects of Modern Persian, Persian with dialectal traits, and closely related languages. In the cities of Yazd and Kermān the Parsis speak the old Gabrī dialect, whereas the Muslims speak Persian. Among other central dialects are Nātanzī, Sōī, Khunsārī, Gazī (near Eṣfahān), Sīvandī (northeast of Shīrāz), Vafsī, and Ashtiyānī, to name but a few.

Semnānī, spoken east of Tehrān, forms a transitional stage between the central dialects and the Caspian dialects. The latter are divided into two groups, Gīlakī and Māzandarānī (Tabarī). Also closely related is Tālishī, spoken on the west coast of the Caspian Sea on both sides of the border with Azerbaijan. To this northwestern group belong the so-called southern Tātī dialects spoken south and southwest of Qazvīn, as well as the scarcely known dialects of Harzan and Galinqaya spoken northwest of Tabriz. The name Tātī is usually applied to the dialects spoken in Russian Dagestan and northeastern Azerbaijan. They differ little from Modern Persian.

Of the several dialects of Fārs province, only Larī, southeast of Shīrāz, is notably distinctive. Kumzarī in Oman and the Lur dialects of the southwest also differ little from Persian.

Kurdish dialects

There are many dialects of Kurdish, the widely spoken West Iranian language that is thought to occupy a dialectal position intermediate between Balochi and Persian. Three main dialect groups can be distinguished—northern, central, and southern. A systematic study has been made of the dialects of Iraq, which include ʿAqrah (Akre), ʿAmādīyah, Dahūk, Shaykhān, and Zākhū in the northern group, and Irbīl (Arbīl), Bingird, Pishdar (Pizhdar), Sulaymānīyah (Suleimaniye), and Wārmāwah in the central group. The Central Mukrī dialect is spoken in the extreme west of Iran, south of Lake Urmia.

Gorānī is spoken in several dialects, mainly in the Zagros Mountains, and it is strongly influenced by the surrounding Kurdish dialects. The Gorānī dialect of Hawrāman, Hawrāmī, is notable for its many archaic features. Closely related to Gorānī is Zaza (Dimli), which is spoken west of Iran.

Historical survey of the Iranian languages. *The Iranian protolanguage and its development.* By the time Iranian begins to be attested in the 6th century BC, the language is already found differentiated into several distinct languages. Scholars have reconstructed the sound system and some of the grammatical features of Common Old Iranian, the protolanguage that preceded these dialects.

The phonological system that underlay Common Old Iranian was by and large maintained everywhere throughout the Iranian-speaking world. It consisted of the following distinctive consonant sounds:

k	g	x	[ɣ]	ŋ
č	ǰ			
p	b	f	[β]	m
t	d	θ	[ð]	n
š	ž			
y	r	l	w	h

Unfamiliar symbols are taken from the International Phonetic Alphabet, or are conventional transcriptions (*e.g.*, *š* for the *sh* sound in "ship," *ž* for the *zh* sound in "azure," *č* for *ch* in "church," and *ǰ* for *j* in "jam"). The voiced fricatives (*i.e.*, the first three consonants represented in the fourth column—ɣ, β, and ð), which are produced with vibrating vocal cords and local friction, may be regarded as variants of the voiced stops (*e.g.*, *g*, *b*, *d*); but they are characteristic of Iranian languages generally and especially of the eastern Iranian languages. In addition to these sounds Old Persian had another sibilant sound, often transcribed as *ç* or *ss*, which developed from the cluster *θr* (pronounced as the *thr* in "three"). In Middle Persian it fell together with the *s* sound. The most noticeable alteration of the old sound system is the introduction in some languages of additional series of consonants under the influence of neighbouring languages. Thus, Ossetic has a series of ejective sounds (uttered with a simultaneous glottal stop) on the pattern of the unrelated Caucasian languages; and a number of Iranian languages have a retroflex series (produced with the tongue tip curled up toward the roof of the mouth) as a result of contact with Indo-Aryan languages.

Changes from Indo-European sounds to Iranian sounds

Some of the differences between Iranian languages arose as a result of different developments of the earlier sounds. Thus, the Indo-European sounds *ḱ*, *ǵ*, and *ǵh* resulted in Indo-Iranian *ś*, *ź*, and *źh*, which in turn became *s*, *z*, and *z*, respectively, in Avestan but *θ*, *d*, and *d* in Old Persian. Hence, Indo-European **ḱm̥tó-* "hundred" became Indo-Iranian **śatá-*, attested by Old Indo-Aryan *śatá-*, and then Avestan *sata-*, but Old Persian *θata-*. Nevertheless, *θ* and *d* as well as *s* and *z* belong to the basic pattern, the difference being merely distributional.

The main source of differentiation is in the variation of consonant cluster development and that of groups of consonants and semivowels. Here again it is mainly a question of distributional differences. Thus, the Indo-European group **ḱu̯* became Indo-Iranian **śu̯*, retained in Old Indo-Aryan in the spelling *śv* of the standard transcription. Indo-Iranian **śu̯* developed variously in Iranian: *s* in Old Persian, *sp* in Avestan and Median, *ś* (written *śś*) in Khotanese, and *š* in Wakhī. These developments can be seen in the following forms of the Indo-European word **eḱu̯o-* "horse": Old Indo-Aryan *áśva-*, Avestan and Median *aspa-*, Old Persian *asa-*, Khotanese *aśśa-*, and Wakhī *yaš*. Yet another development can be seen in Ossetic, in which the word for "mare," Avestan *aspā-*, appears as Digor *äfsä* and Iron *yäfs*.

The vowel system of Common Old Iranian consisted of short and long varieties of *a*, *i*, and *u*, and a neutral vowel *ə* (similar to the *a* in "sofa"). This analysis assumes that the Indo-Iranian vocalic *r* (*r̥*) had already developed to *ər* in Proto-Iranian, just as its long counterpart became *ar*. An early and general monophthongization of the diphthongs *ai* and *au* to *ē* and *ō*, respectively, also must be considered characteristic, although it should not be ascribed to Common Old Iranian as is sometimes done. This basic system was almost everywhere maintained, sometimes with the addition of one or two distinctive vowel sounds (phonemes).

The Old Iranian stage. Old Persian was the language of the Achaemenid court. It is first attested in the inscriptions of Darius I (ruled 522–486 BC), of which the longest, earliest, and most important is that of Bīsitūn. At Bīsitūn are also inscribed versions of the same text in Elamite and Babylonian, and fragments of an Aramaic version on papyrus documents from Elephantine (modern Jazīrat Aswān) also exist. Old Persian words and names also are to be found in large numbers as loanwords in contemporary Elamite sources and in 5th-century-BC Aramaic documents.

As early as the time of Darius the Great's successor, Xerxes I (ruled 486–465 BC), the inscriptions show linguistic tendencies characteristic of the development from Old to Middle Persian. After Xerxes the production of original Old Persian inscriptions declined, probably as a result of the wider adoption of Aramaic and Elamite as the usual means of writing. With Artaxerxes III (ruled 359/358–338 BC), Old Persian inscriptions came to an end. The break is marked by Alexander's destruction of Persepolis in 330 BC.

Alternate terms for Avestan

By far the largest part of attested Old Iranian is written in the language now usually called Avestan, after the Avesta, the name given to the collection of works forming the scripture of the Zoroastrians. The name itself is Middle Persian. In former times this language was called Zend, another Middle Persian word, which refers to the Middle Persian (Pahlavi) commentary on the Avesta. Because the homeland of the Avestan language was long thought to be in Bactria, it was often in the past called Bactrian. Bactrian is now used to designate a different Iranian language belonging to the Middle Iranian period.

Since the beginning of the 20th century it has been generally accepted that the homeland of the Avesta was Khwārezm, which in ancient times included both Merv and Herāt. Merv is now in Turkmenistan, Herāt in northwestern Afghanistan.

The oldest part of the Avesta is known as the *Gāthā*s, the poems composed by Zoroaster (Zarathustra), the founder of the Zoroastrian religion. His date is uncertain but is traditionally ascribed to the 7th to 6th century BC. The so-called *Khurda Avesta* ("Little Avesta") is a miscellany of texts of later date, the oldest parts of which may have been composed about 400 BC. The language of the *Khurda Avesta* is different in many details from that of the more archaic language of the *Gāthā*s, and it may even represent a different dialect. Many uncertainties surround the detailed interpretation of the Avesta as a result of the method of transmission. The Avesta was not recorded until after the language had ceased to be used, except by Zoroastrian priests. The present manuscripts date from the 13th century and later, although they reflect the recording of the priestly tradition in the special Avestan script during the 6th century AD.

The Middle Iranian stage. Middle Persian, the major form of which is called Pahlavi, was the official language of the Sāsānians (AD 224–651). The most important of the Middle Persian inscriptions is that of Shāpūr I (d. AD 272), which has parallel versions in Parthian and Greek. Middle Persian was also the language of the Manichaean and Zoroastrian books written during the 3rd to the 10th century AD.

The extant literature of the Zoroastrian books is much more extensive than that of the Manichaean texts, but the latter have the advantage of having been recorded in a clear and unambiguous script. Moreover, the Middle Persian of the Zoroastrian books does not simply represent the spoken language of the writers of the 9th-century Zoroastrian texts. It is probable that they spoke early Modern Persian and that their speech often impinged upon their writing but that they strove to write the Middle Persian of several centuries earlier as it was attested in the inscriptions of the early Sāsānian dynasty when Middle Persian was the koine. By contrast, in the case of Manichaean Middle Persian, some texts survive unchanged from the 3rd century AD, the time of the Persian teacher Mani himself (AD 216–274).

Records of Parthian

Very little Parthian survives from the pre-Sāsānian period. A large number of Parthian ostraca (inscribed pottery fragments) from the 1st century BC were discovered at Nisa near modern Ashkhabad, but they are inscribed in ideographic Aramaic (*i.e.,* Aramaic writing that uses Aramaic words as symbols to represent Parthian words). Dating before the 3rd century are a document from Hawrāman, some coin legends, and a dated grave stele.

The most copious and important material in Parthian is the work of the Sāsānian kings of the 3rd century, who added a Parthian version to their inscriptions—Ḥājjīābād, Naqsh-e Rustam (Ka'be yi Zardusht), and Paikūla. A few decades later Parthian disappeared as a result of the rise of the Sāsānians and the predominance of their native tongue, Middle Persian. Manichaean Parthian of the 3rd century was preserved as a church language in Central Asia.

The oldest surviving Sogdian documents are the so-called Ancient Letters found in a watchtower on the Chinese Great Wall, west of Tun-huang, and dated at the beginning of the 4th century AD. Most of the religious literature written in Sogdian dates from the 9th and 10th centuries. The Manichaean, Buddhist, and Christian Sogdian texts come mainly from small communities of Sogdians in the T'u-lu-p'an (Turfan) oasis and in Tun-huang. From Sogdiana itself there is only a small collection of documents from Mt. Mugh in the Zarafshān region, mainly the business correspondence of a minor Sogdian king, Dewashtich, from the time of the Arab conquest about 700.

The relationship of the various forms of Sogdian to one another has not yet been sufficiently investigated, so that it is not clear whether different dialects are represented by the extant material or whether the differences can be accounted for by reference to other relevant factors, such as differences of script, period, subject, style, or social milieu. The importance of social milieu can be seen by comparing the elegant Manichaean literature directed to the court with the more vulgar language of the Christian literature directed to the lower classes.

Of the Saka dialect known as Tumshuq very little has survived, and despite its evidently close relationship to the much better known Khotanese dialect, full interpretation has proved difficult. Knowledge of Khotanese is more firmly based on a substantial corpus of material, including extensive bilingual texts. Although the chronological range of the extant Khotanese material is limited to only a few centuries, probably the 7th to the 10th, a rapid development of the language is apparent. At the phonological level, most noticeable is the loss of syllables between the older and later stages of the language. Thus, *hvatana-* "Khotanese" at the oldest stage is successively weakened to *hvatäna-, hvaṃna-, hvana-, hvaṃ.* At the morphological level, most striking is the tendency to simplify the case endings and even to replace them by analytical expressions, constructions of two or more words. Thus, Late Khotanese has *rakṣaysā hīya rāde* "kings of the *rākṣasas,*" whereas Old Khotanese would have *rakṣaysänu rrunde.* The Old Khotanese *-änu* ending is unmistakably genitive plural, but the Late Khotanese *-ā* is merely a general oblique plural ending and has been reinforced by *hīya* "own," used to mean "of."

Buddhist influence on Khotanese

Khotan was a great centre of Buddhism during the 1st millennium AD, and all the surviving literature in Khotanese is either Buddhist or coloured by Buddhism. Even in business documents and official letters the Buddhist background is usually not difficult to discern. It can scarcely be coincidental that the Buddhist literature of Khotan, flourishing so vigorously during the 10th century, ended abruptly with the Muslim conquest at the beginning of the 11th.

Little survives of Bactrian and Scytho-Sarmatian. Knowledge of Bactrian is based almost entirely on a single inscription of 25 lines from Āteshkadeh-ye Sorkh Kowtal in northern Afghanistan. Even less is known of Scytho-Sarmatian.

Little is also known of Old Khwārezmian; that is, Khwārezmian written in the indigenous Khwārezmian script. Apart from a few coin legends and inscriptions on silver vessels, the material that survives consists of inscriptions of the 2nd century AD from Topraq-qal'ah (Toprakkala) and of the 7th from Toqqal'ah, archaeological sites in Uzbekistan. Much more is known of Late Khwārezmian, written in the Arabic script. This material is found mainly in two Arabic works, the 13th-century *fiqh* work of Mukhtār az-Zāhidī, called the *Qunyat almunyah,* and the Arabic dictionary *Muqaddimat al-Adab* of az-Zamakhsharī (1075–1144), of which a manuscript glossed in Khwārezmian was found.

Modern Iranian. Of the modern Iranian languages, by far the most widely spoken is Persian, which, as already indicated, developed from Middle Persian and Parthian (with elements from other Iranian languages such as Sogdian) as early as the 9th century AD. Since then, it has changed little except for acquiring an increasing proportion of loanwords, mainly from Arabic. Persian has been a literary language since the 9th century, and there is an increasing awareness of the continuity of its literary tradition with the earlier periods.

As the national language of Iran in succession to Middle Persian, it has for centuries strongly influenced the other Iranian languages, especially on Iranian territory. In fact, it seems likely that, with the increase of modern methods of communication, Persian will eventually supplant entirely most of the other languages and dialects. Against this trend stand only Kurdish and Balochi, the speakers of which tend to regard their languages as an expression of their particular identities. Nevertheless, even Kurdish and Balochi have been and continue to be strongly influenced by Persian.

Modern Iranian languages

Outside Iran the situation is rather different. In Afghanistan the first national language is Pashto, even though Persian is the official second language. Pashto became the official language by royal decree in 1936, and literary activity has been encouraged by the Pashto Ṭolana (Pashto Society) of Kabul. During the Soviet period both Ossetic and Tajik received official encouragement; nevertheless, both languages were displaced by the Russian language as the language of administration.

Other languages also compete with Ossetic and Tajik. Though it has a large body of folk epics, Ossetic became a literary language only in the second half of the 19th century. By contrast, the neighbouring Georgian has a still flourishing ancient literary tradition dating back to the 5th century AD and has many more speakers. Tajik, on the other hand, has a lifeline through its close connection with Persian, but it too has been retreating before Uzbek, an unrelated language of the Turkic group.

Characteristics of the Iranian languages. All Iranian languages show in their basic elements the characteristic features of an Indo-European language. Apart from the extensive borrowing of Arabic words in Modern Persian, the Iranian languages have scarcely been affected by unrelated languages, with the notable exception of Ossetic, which has been strongly influenced by the neighbouring Caucasian languages. Some dialects of Tajik have been very receptive to Uzbek elements. In the case of languages in contact with Indian civilization, the most noticeable non-Iranian feature often taken over is the Indo-Aryan series of retroflex sounds. These are foreign to Indo-Aryan itself, being a result of the influence of the Dravidian languages.

The elaborate phonological and morphological structure of the Indo-European parent language has been progressively simplified in the development of the Iranian languages. The basic phonological structure of Common Old Iranian has on the whole been maintained, but the morphological system has continued to be simplified. There has been a constant move in almost all Iranian languages toward an analytic structure; *i.e.*, the use of prepositions and word order rather than case endings to indicate grammatical relationships.

Development of Iranian fricative sounds

Phonology. The most characteristic features of the Iranian phonological system are those that distinguish it from the Indo-Aryan system. These are the development of various fricative sounds (indicated in phonetic symbols as x, f, *θ*, and later ɤ, β, ð), and of the voiced sibilant sounds *z* and *ž*. Even in Iranian, however, these sounds did not persist universally. In western Middle Iranian the *θ* sound was lost, and it is rare in the modern languages. In Pashto the inherited *f* sound has been discarded. Baluchi, except in the extreme east, is entirely without fricatives. Voiced bilabial and dental fricative sounds (β and ð) were recorded in some early manuscripts of Modern Persian, but they became *b* and *d* by the 13th century

Two negative features have also resulted in differentiation between Indo-Aryan and Iranian. One is the result of the coalescence in Proto-Iranian of aspirated and unaspirated voiced stops. Thus, Indo-European **b* and **bh* were maintained in contrast in Indo-Aryan as *b* and *bh,* but they fell together in Iranian as *b.* This resulted in an alteration of the phonological structure because the number of consonant contrasts (oppositions) was reduced. The other negative feature is the absence of the retroflex consonants from Iranian except as a later importation in contiguous regions.

Other divergences in development, such as the change of an *s* sound to *h* in Iranian, brought about a difference in distribution rather than in structure because *h* developed also in Indo-Aryan but from Indo-Iranian **źh* and **gh* before front vowels (*e.g.*, *e* and *i*). The features discussed here are illustrated in Table 6.

In Old Iranian the stress lay on the next to the last syllable if it was heavy (*i.e.,* contained a long vowel or was closed by a consonant)—otherwise on the preceding syllable. With the loss of final unstressed vowels in the development of many Iranian languages, the stress often came to be on the final syllable. End stress is characteristic of Modern Persian.

Simplification of the case system

Grammar. In Old Persian the Indo-European inflectional system appears considerably simplified. In particular, the genitive and the dative coalesced into one case and the instrumental and ablative into another. Moreover, in the plural the nominative and accusative cases are not distinguished. This reduced system is still found in the Middle Iranian period in Old Khotanese and to a certain extent in Sogdian. Eastern Iranian is in this respect more conservative than western. By the Middle Iranian period, western Iranian had abandoned nominal (noun, adjective, pronoun) inflection altogether, as is the case with Middle and Modern Persian and with Parthian. In some languages, both western and eastern, two or, rarely, three cases survive. Ossetic is quite exceptional in maintaining an elaborate case system; it is partly a result of secondary, purely Ossetic developments.

Table 6: Phonetic Developments in Indo-Iranian Languages
key: NP—New Persian; Bal.—Baluchi; Yaghn.—Yaghnobi

Sanskrit	Avestan	Old Persian	modern Iranian	English translation
kratu-	xratu-	xratu-	NP xirad	"insight"
viś-	vis-	viθ-	Bal. gis	"house"
jānáti ("he knows")	zān(ā)	dān(ā)-	NP dān-	"know"
bandh-	band-	ba(n)d-	NP band-	"bind"
bhūta-	būta-		NP būd	"been"
sacā ("with, at the same time as. . .")	hačā	hačā	NP az	"from"
han-	ǰan-	ǰan-	Bal. ǰan-	"strike"
abhra-	aβra-		NP abr	"cloud"
mṛga- ("deer")	mərəɣa-		NP murɣ	"bird"
nir-ay-		nij-ay-*	Yaghn. nīž-	"go out"
pramāṇa- ("measure, authority")		framānā-	NP farmān	"command"
sthūṇā-	stunā-	stūnā-	NP sitūn	"pillar"

*In Old Persian *nij-ay-* "go out," *j* is written for *ž*, which was not represented in the script.

The elaborate conjugational system of the Indo-European verb followed a similar path to disintegration. In particular, the whole past tense system was given up by the Middle Iranian period. Only a few relics remain of the Indo-European system, such as the partial survival of the augment (a prefixed vowel or lengthening of the initial vowel) in the Sogdian imperfect tense. But a new past tense system developed, based on the old past participle, often combined with auxiliary verbs. Many languages distinguish between transitive and intransitive verbs in the past tense system; and in some, such as Khotanese and Pashto, even gender and number are distinguished.

The present tense system was far better preserved. The dual number was in retreat in Old Iranian and is not attested later. The middle voice, a form that indicates that a person or thing both performs and is affected by the action represented, was generally abandoned by the Middle Iranian period, although middle voice inflection is well represented in Khotanese. With these qualifications, the endings of the present indicative (active) have been generally well preserved. A variety of imperative, subjunctive, and optative forms, partly based on inherited forms and partly the result of innovation, is found especially in the eastern languages, including Ossetic.

Rigidity of word order is, on the whole, most characteristic of those languages, such as Persian, that have gone furthest in the reduction of the inherited morphological system.

Vocabulary. The Islāmic conquest of Iran during the 7th century entailed not only a change of religion but also a change of language. The sacred language of Islām was Arabic, and the proportion of Arabic words used in Persian rapidly increased until it reached something like the 40 to 50 percent of the present day. Before the introduction of the Arabic element, most loanwords were mainly from other Iranian languages. Most familiar is the extensive borrowing from Median found in Old Persian. In later periods, Modern Persian borrowed words extensively from Turkish and from European languages. Persian is itself the donor language in the case of the other Iranian languages, all of which have drawn upon its vocabulary.

Loanwords from languages of India

Buddhism was similarly responsible for the large proportion of Indo-Aryan words, both Sanskrit and Prākrit, found in Sogdian and especially in Khotanese. A considerable Indian element occurs in the vocabulary of those modern Iranian languages that have been or are in contact with modern Indo-Aryan languages in the northwest, such as Lahnda and Sindhi. There the Dardic languages have also been influential. Baluchi has also borrowed from Brahui, a Dravidian language spoken in Baluchistan in Pakistan.

Ossetic occupies an exceptional position. Most of its Persian and Arabic borrowings have come to it through Turkish, but more striking are the large number of words borrowed from the Caucasian languages, especially Georgian. In modern times, Ossetic continues to be influenced by Russian.

Writing systems. Iranian languages have been written in many different scripts during their long history, although various forms of Aramaic script have been predominant. Modern Persian is written in Arabic script, which is of Aramaic origin. For writing the Persian sounds *p, č, ž,* and *g,* four letters have been added by means of diacritical marks. By the addition of further letters, this Perso-Arabic script has been adapted to write not only the other main modern Iranian languages, Pashto, Kurdish, and Baluchi, but also those minor ones that are occasionally recorded. An advantage of the use of this consonantal script is that

Table 7: The Persian Alphabet

consonants					equivalents		approximate pronunciation
alone	initial	medial	final	name	EB preferred	alternatives	
ا	ا	ـا	ـا	alef	*		*
ب	بـ	ـبـ	ـب	be	b		*baby*
پ	پـ	ـپـ	ـپ	pe	p		*pepper*
ت	تـ	ـتـ	ـت	te	t		*tie*
ث	ثـ	ـثـ	ـث	še	s̱	s, th	*sand*
ج	جـ	ـجـ	ـج	jīm	j	dj	*job*
چ	چـ	ـچـ	ـچ	che	ch	č	*chin*
ح	حـ	ـحـ	ـح	he hoti	ḥ	ḥ	*hat*
خ	خـ	ـخـ	ـخ	khe	kh	kh	Ger. Bu*ch*
د	د	ـد	ـد	dāl	d		*did*
ذ	ذ	ـذ	ـذ	ẕāl	ẕ	z, dh	*zone*
ر	ر	ـر	ـر	re	r		*rip*
ز	ز	ـز	ـز	ze	z		*zone*
ژ	ژ	ـژ	ـژ	zhe	zh	zh	a*z*ure
س	سـ	ـسـ	ـس	sīn	s		*sand*
ش	شـ	ـشـ	ـش	shīn	sh	sh	*sh*y
ص	صـ	ـصـ	ـص	ṣād	ṣ	ṣ	*sand*
ض	ضـ	ـضـ	ـض	ẕād	ẕ	z	*zone*
ط	طـ	ـطـ	ـط	ṭā	ṭ	ṭ	*time*
ظ	ظـ	ـظـ	ـظ	ẓā	ẓ	ẓ	*zone*
ع	عـ	ـعـ	ـع	ʿeyn	ʿ		†
غ	غـ	ـغـ	ـغ	gheyn	gh	gh, q	‡
ف	فـ	ـفـ	ـف	fe	f	fe	*fif*ty
ق	قـ	ـقـ	ـق	qāf	q	k	‡
ک	کـ	ـکـ	ـک	kāf	k		*k*in
گ	گـ	ـگـ	ـگ	gāf	g		*g*o
ل	لـ	ـلـ	ـل	lām	l		*lily*
م	مـ	ـمـ	ـم	mīm	m		*maim*
ن	نـ	ـنـ	ـن	nūn	n		*n*o
و	و	ـو	ـو	vāv	v	w	*v*an§
ه	هـ	ـهـ	ـه	he havaz	h		‖
ی	یـ	ـیـ	ـی	ye	y		*y*et¶

vowels, diphthongs, and special diacritical marks		equivalents		approximate pronunciation
letter or sign	name	EB preferred	alternatives	
آ	alef maddeh	ā	á	*a*rm
ىٰ	alef maqṣūreh	ā	á	*a*rm
ا	alef	*		*
و	vāv	ū		f*oo*d§
ی	ye	ī		bl*ee*d¶
َ	fatḥeh (or zebar)	a		m*a*p
ِ	kasreh (or zīr)	e		b*e*t
ُ	ẕammeh (or pīsh)	o		b*o*ne or *o*range
ـِی-(ـِی)	kasreh ye	ī		bl*ee*d
ـُو-(ـُو)	ẕammeh vāv	ū		f*oo*d
ـَی-(ـَی)	kasreh ye sāken	ey	ay, ai	f*a*de
ـَو-(ـَو)	ẕammeh vāv sāken	ow	aw, au	b*o*ne
ـْ	sokūn (or jazm)	omit		♀
ـة	tā marbūṭah	eh, ah, or at		♂
ـّ	tashdīd	double consonant		me*dd*ling, etc.
ء	hamzeh	initial, omit; medial and final,’		□
ـِ ,ـی —	eẕāfeh	-e, -ye	-i, -yi	◇

*Initially *a* or *e*, pronounced *m*a*p* or *b*e*t*; medially and finally, *ā*, pronounced *a*rm. †A glottal stop, as in New York or Cockney "bottle." ‡A guttural *gh*; also medially and finally often softened as in French *r*ien. §As a consonant, *v*an; as a vowel, f*oo*d. ‖Generally silent in final position; otherwise *h*at. ¶As a consonant, *y*et; as a vowel, bl*ee*d. ♀Used to show that a consonant is not vocalized. ♂Arabic feminine ending usually not pronounced in Persian except as silent *h*. □A pause between two vowels, as in English "di-et," "qui-et." ◇A particle linking a qualifying noun or adjective with a noun; usually not written. Transliterate "-e" (with hyphen) after final consonants (except silent *h*), "-ye" (with hyphen) after silent *h* and vowels.

by not defining vowel qualities it is possible to include local dialect variations to a considerable extent.

Two modern Iranian languages spoken on Soviet territory are currently written in a modified version of the Russian alphabet: Tadzhik and Ossetic. Soviet scholars have, however, tended to use modified Latin alphabets to record the minor languages that have no literary tradition, such as some of the Pamir languages. Ossetic has also been written in the Georgian script.

Old Persian was written with a cuneiform syllabary, the origin of which is still hotly disputed. Middle Persian, Parthian, Sogdian, and Old Khwārezmian were recorded in various forms of Aramaic script. Two forms of this script as they developed for writing Sogdian were adopted by the Uighurs. In its cursive form this script spread even further, to the Mongols and Manchus. Three other scripts are important for the remaining Middle Iranian languages: Greek script for Bactrian, Arabic script for Late Khwārezmian, and varieties of Central Asian Brāhmī script of Indian origin for Khotanese and Tumshuq.

The Aramaic script was not systematically adapted to the writing of Middle Iranian; and despite the introduction of a variety of diacritical marks to differentiate letters, considerable ambiguity remained. Moreover, several letters tended to coalesce in form. In this respect, the Pahlavi script, used for writing the Middle Persian of the Zoroastrian books, developed furthest. In it, the original 22 letters of the Aramaic alphabet have been reduced to 14, which are further confused by the use of numerous ligatures (linked letters). It was the realization that this script was inadequate to record precisely the traditional pronunciation of the sacred text of the Avesta that led the Zoroastrian priests to devise the elaborate Avestan script, which, with its 48 distinct letters formed by differentiation out of the 14 used for Pahlavi, was well suited to the task. (R.E.E.)

Greek language

Greek is an Indo-European language whose history can be followed from the 14th century BC to the present day. Its documents cover 34 centuries, a longer period of time than that of any other Indo-European language. There is an Ancient phase, subdivided into a Mycenaean period (texts in syllabic script attested from the 14th to the 13th century BC) and Archaic and Classical periods (beginning with the adoption of the alphabet, from the 8th to the 4th century BC); a Hellenistic and Roman phase (4th century BC to 4th century AD); a Byzantine phase (5th to 15th century AD); and a Modern phase.

Separate transliteration tables for Classical and Modern Greek accompany this article. Some differences in transliteration result from changes in pronunciation of the Greek language; others reflect convention, as for example the χ (*chi* or *khi*), which was transliterated by the Romans as *ch* (because they lacked the letter *k* in their usual alphabet). In Modern Greek, however, the standard transliteration for χ is *kh*. Another difference is the representation of β (*bēta* or *víta*); in Classical Greek it is transliterated as *b* in every instance, and in Modern Greek as *v*. The pronunciation of Ancient Greek vowels is indicated by the transliteration used by the Romans. Y (*upsilon*) was written as *y* by the Romans, indicating that the sound was not identical to the sound of their letter *i*. Modern Greek υ (*ípsilon*) is transliterated as *i*, indicating that the sound used today differs from that of the ancient υ. (See Table 8 for the Classical Greek alphabet and numerals and Table 9 for the Modern Greek alphabet.)

GENERAL CONSIDERATIONS

In the first half of the 2nd millennium BC, the "Proto-Greeks" (Indo-European ancestors of the Greeks) established themselves on the Greek peninsula, where their language developed into Greek. Later, Greek-speaking people occupied most of the islands of the Aegean and, about 1000 BC, the west coast of Anatolia. With few exceptions that is still the area occupied by the Greek language today. In the second quarter of the 1st millennium BC, a vast "colonial" movement took place, resulting in establishments founded by various Greek cities all around the Mediterranean and the Black Sea, especially in southern Italy and Sicily. This extension of the linguistic area of Greek lasted only a few centuries; in the Roman period, Latin, more or less rapidly, took the place of Greek in most of these ancient colonies. After the conquest of Asia Minor, Syria, and Egypt by Alexander the Great, Greek was the standard language of the rulers in the new urban centres of these countries until the invasions of the Arabs and the Turks. "Colonial" Greek survived longest at Byzantium, as the official language of the Eastern Empire.

Relationship of Greek to Indo-European. Ancient Greek is, next to Hittite, the Indo-European language with documents going furthest back into the past. By the time it emerged in the second half of the 2nd millennium BC, it had already acquired a completely distinct character from the parent Indo-European language. Its linguistic features place it in a central region on the dialect map that can be reconstructed for Proto-Indo-European; the ancient languages with which it has the most features in common are little-known ones such as Phrygian. In the study of Indo-European dialectology, phonetic data are the most readily available and provide the most information. In this respect the position of Ancient Greek is as follows.

Distinctive traits of Ancient Greek

The vowels of *a* and *o* quality, both short and long, remain distinct, whereas they are completely or partially confused in Hittite, Indo-Iranian, Baltic, Slavic, and Germanic. Greek is the only language that distinguishes by three different qualities (*ĕ*, *ă*, *ŏ*) the secondary short vowels resulting in certain positions from the three laryngeal sounds, $*H_1$, $*H_2$, $*H_3$, of Indo-European. (An asterisk preceding a sound or word indicates that it is not an attested, but a reconstructed, hypothetical form. For a discussion of these laryngeal sounds, see *Indo-European languages*.) Greek keeps the distinction between the original voiced stops and voiced aspirated stops of Indo-European (*e.g.,* Indo-European **d* becomes Greek *d,* and Indo-European **dh* becomes Greek *th*), whereas Iranian, Slavic, Baltic, and Celtic confuse them. (Some linguists, however, assume that Greek *th* continues Indo-European *th* and that Greek *d* goes back to an Indo-European glottalized stop.) Greek avoids the general shifts of stop consonants that are displayed, independently, by Armenian and Germanic, as well as the change of palatal stops (*k,* etc.) into affricates (*ts,* etc.) or spirants (*s,* etc.) in Indo-Iranian, Armenian, Baltic, and Slavic. In these respects Ancient Greek is conservative, as are, generally speaking, the western Indo-European languages (Italic and Celtic). On the other hand, it does show innovations. One of these, the devoicing of the original voiced aspirated stops, is shared with Italic, although it is realized in different ways (**dh-* yields Greek *th-*, Latin *f-*, Osco-Umbrian *f-*); but others are foreign to Italic. The latter include, for example, the weakening of spirants and semivowels at the beginning of words before a vowel, the evolution of **s-* to *h-* (pre-Mycenaean), and **y-* to *h-* (contemporary with Mycenaean).

Morphological criteria must, of course, be taken into account in defining the position of a language. It should be noted that there are few grammatical innovations shared by Greek and Italic, apart from the extension to nouns of the pronominal ending of the genitive feminine plural **-āsōm* (Greek *-āōn;* Latin *-ārum,* Umbrian *-aru,* Oscan *-azum*) and of the pronominal ending of the nominative masculine plural **-oi* (Greek *-oi;* Latin *-ī*). The last innovation, however, is not shared with Osco-Umbrian but is found instead in Germanic (in the strong declension of adjectives) and partly in Celtic. The dialectal individuality of Greek is very clearly marked in the organization of the verb (see below), which is without parallel except for an approximation in Indo-Iranian.

Greek syllabaries. Starting from a foreign script known as Linear A (used in Crete to record a native language known as Minoan), the Greeks devised, toward 1500 BC at the latest, a syllabic script to record their own language. Known as Linear B, this script was deciphered in 1952 by the British architect Michael Ventris and the British classicist and linguist John Chadwick. At present more than 100 very short Linear B inscriptions painted on vases have been found in Crete and in continental Greece

Table 8: Classical Greek Alphabet and Numerals

letters: capital	letters: lower-case	letters: combi-nations	name	equivalent: EB preferred	equivalent: alter-natives	approximate classical Attic pronunciation
Α	α, α*		alpha	a		*a*re
		αι		ae in proper nouns, ai in common words	e	*i*ce
		αυ		au		n*ow*
Β	β		beta	b		*b*aby
Γ	γ		gamma	g		*g*o
		γγ		ng		a*ng*le
		γκ		nk	nc	i*nk*
		γξ		nx		tha*nks*
		γχ		nch	nkh	i*n c*ase
Δ	δ, ∂*		delta	d		*d*og
Ε	ε		epsilon	e		b*e*t
		ει		ei	e or i	d*ay*
		ευ		eu		b*e*t + n*ow*
Ζ	ζ		zeta	z		u*sed*
Η	η		eta	ē	e	*ai*r
		ηυ		ēu	eu	*airw*ay
Θ	θ, ϑ*		theta	th		*t*in
Ι	ι		iota	i		*e*ven or p*i*n
Κ	κ		kappa	c in proper nouns, k in common words		po*ck*et
Λ	λ		lambda	l		*l*i*l*y
Μ	μ		mu	m		*m*ai*m*
Ν	ν		nu	n		*n*ot
Ξ	ξ		xi	x		a*x*
Ο	ο		omicron	o		Ger. s*o*
		οι		oe in proper nouns, oi in common words	oe	Ger. s*o* + d*ay*
		ου		ou		*ow*n
Π	π		pi	p		s*p*in
Ρ	ρ		rho	initial, rh; medial, r		*r*ose
		ρρ		rrh		Ger. Natu*rr*echt
Σ	σ†		sigma	s		*s*and
Τ	τ		tau	t		s*t*ay
Υ	υ		upsilon	y	u	Fr. d*u*
		υι		ui		Fr. concl*ui*ez
Φ	ɸ, φ*		phi	ph		*p*in
Χ	χ		chi	ch	kh	*k*in
Ψ	ψ		psi	ps		perha*ps*
Ω	ω		omega	ō	o	c*a*ll

numerals

Greek	Arabic	Greek	Arabic	Greek	Arabic
α′	1	ιε′	15	ο′	70
β′	2	ις′	16	π′	80
γ′	3	ιζ′	17	ϙ′	90
δ′	4	ιη′	18	ρ′	100
ε′	5	ιθ′	19	σ′	200
ς′	6	κ′	20	τ′	300
ζ′	7	κα′	21	υ′	400
η′	8	κβ′	22	ɸ′	500
θ′	9	κγ′	23	χ′	600
ι′	10	κδ′	24	ψ′	700
ια′	11	λ′	30	ω′	800
ιβ′	12	μ′	40	ϡ′	900
ιγ′	13	ν′	50	͵α	1,000
ιδ′	14	ξ′	60		

*Old-style character. †Final, ς

(*e.g.,* Thebes), where they were imported from Crete. The major source of Linear B inscriptions are some 4,500 unbaked clay tablets found at Knossos (1400–1350 BC—this date has been questioned), and at Thebes, Tiryns, Mycenae, Pylos, and Chania (1250–1200 BC). There are no literary texts and hardly any continuous texts (only a small number of complete sentences exist); the tablets contain accounts of the great Mycenaean palaces and their dependencies, written in the Greek language, in a very concise style.

Signs of Linear B

The Linear B syllabary consists of about 90 signs. In principle, each sign represents a syllable beginning with one consonant and ending with a vowel. Thus there are five different signs for *ta, te, ti, to, tu,* but there is no sign for the consonant *t* without a following vowel. As an initial syllable may be formed by just a vowel, there are also signs for *a, e, i, o, u.* The script does not distinguish *r* and *l,* unvoiced and voiced consonants (except for /d-/), nonaspirated and aspirated consonants; so the sign *pa* can be read in Greek as *pa, ba,* or *pha.* Final consonants are omitted, and consonants followed by other consonants are either omitted or expressed by means of the sign corresponding to the next vowel (*e.g., pe-ma* for *sperma, ta-to-mo* for *stathmos*). Consequently, the spellings are often ambiguous, such as *pa-te* for *pantes* and *patēr, pa-si* for *pansi* and *phāsi.* This inconvenient script and the nature of the documents make Mycenaean inscriptions harder to exploit and less rich in data than the later alphabetic inscriptions; but the information that can be gathered on the state of Greek five centuries before Homer, incomplete as it may be, is of capital importance.

Another syllabary, distantly related to Linear B, was in use in Cyprus. From the 11th to the 3rd century BC it was used to record a native language of the island (Eteocypriot) as well as Greek.

The Greek alphabet. The Mycenaean script dropped out of use in the 12th century when the Mycenaean palaces were destroyed, perhaps in connection with the Dorian invasions. For a few centuries the Greeks seem to have been illiterate.

In the 8th century at the latest, but probably much earlier, the Greeks borrowed their alphabet from the Phoenicians in the framework of their commercial contacts. The Phoenician alphabet had separate signs for the Semitic consonants, but the vowels were left unexpressed. The list of Semitic consonants was adapted to the needs of Greek phonology, but the major innovation was the use of five letters with the value of vowels—α (*a*), ε (*e*), ι (*i*), ο (*o*), υ (*u*). The earliest datable inscriptions, both from approximately 725 BC, come from Athens (the Dipylon vase) and the colony of Ischia in the Tyrrhenian Sea (the so-called Nestor's cup).

During the period from the 8th to the 5th century BC, there were local differences in the forms of the letters and in their values from one city to another. Moreover, the primitive Greek alphabet underwent various reforms—the creation of new letters or a new use of old letters, first φ (*ph*), χ (*ch*), then ξ (*ks*), ψ (*ps*), η (*ē*), and ω (*ō*). About 400 BC the alphabet became normalized throughout the Greek world as the Asiatic Ionic form was uniformly adopted.

Greek alphabetic inscriptions

Greek alphabetic inscriptions are numbered in tens of thousands: dedications, epitaphs, decrees, laws, treaties, religious rules, judicial decisions, and so forth. The majority are of Hellenistic or Roman date. The less numerous Archaic inscriptions (8th–5th centuries BC) are of particular interest for their contribution to the knowledge of the dialects (see below). It is only in Hellenistic papyri, and later in Byzantine manuscripts, that the great works of ancient literature (the originals of which have disappeared) are available in the form of copies, some of which are far removed from the originals.

The Greek alphabet, still in use today in Greece in the form it reached during the Hellenistic period, has enjoyed an extraordinary success as a direct or indirect model for other alphabets (notably the Latin alphabet); on it are based the writing systems employed in a great part of the modern world.

ANCIENT GREEK

History and development. From the end of the 4th century BC onward, in the Hellenistic period, Greek gradually obtained a high degree of unity throughout the area it covered (see below *Koine*). In the preceding 10 centuries

Table 9: Modern Greek Alphabet

Greek letters: capital	lower case	combinations	name	equivalents	approximate pronunciation
Α	α, α*		álfa	a	bother
		αι		e	bed
		αϊ		ai	life
		αυ		av/af	lava†, waft
		αϋ		ai	life
Β	β		víta	v	van
Γ	γ		ghámma	gh before α, ο, ου, ω and consonants other than γ, ξ, and χ; y before αι, ε, ει, η, ι, οι, υ, and υι; n before γ, ξ and χ	wit, yet, sing
		γκ		initial, g; medial, ng	go, finger
Δ	δ, ∂*		dhélta	dh; d between ν and ρ	then wondrous
Ε	ε		épsilon	e	bet
		ει		i	even
		εϊ		eï	day
		ευ		ev/ef	revel, left
Ζ	ζ		zíta	z	zone
Η	η		íta	i	fig
		ηυ		iv/if	even, leaf
Θ	θ, ϑ*		thíta	th	thin
Ι	ι		ióta	i	even
Κ	κ		káppa	k	kin, cook
Λ	λ		lámbdha	l	lily
Μ	μ		mi	m	maim
		μπ		initial, b; medial, mb	baby, ambush
Ν	ν		ni	n	not
		ντ		initial, d; medial, nd	dog, fender
		ντζ		ntz	chintz
Ξ	ξ		xi	x	ax
Ο	ο		ómikron	o	saw
		οι		i	even
		οϊ		oi	boy
		ου		u	food
Π	π		pi	p	pin
Ρ	ρ		ro	r	rose
Σ	σ‡		sígma	s	sand
Τ	τ		taf	t	tie
Υ	υ		ípsilon	i initially and between consonants	even
		υι		i	even
Φ	Φ, φ*		fi	f	fifty
Χ	χ		khi	kh	Ger. Buch
Ψ	ψ		psi	ps	perhaps
Ω	ω		oméga	o	bone

*Old-style character. †Pronounced with long *a*. ‡Final, ς.

there had been numerous Greek dialects, which differed in phonetic and morphological details but which were mutually intelligible. The features shared by the local speech of different regions allow the delineation of dialect groups, of which the Greeks themselves were aware. The classifications of modern scholars modify in various ways the classifications made by the ancients but still retain these as their basis. Among the dialects there are a West group, an Aeolic group, an Arcado-Cypriot group, and an Ionic-Attic group. Modern scholars have tried in various ways to combine some of these groups—for example, by considering Aeolic and Arcado-Cypriot as varieties of "central" Greek or by considering Arcado-Cypriot and Ionic-Attic as varieties of "southern" Greek and West Greek and Aeolic as varieties of "northern" Greek.

In regard to the dialects, two very different situations must be distinguished: that established for the period between the 14th and the 12th centuries BC and that for the period between the 8th and the 4th centuries BC.

In Mycenaean times the carriers of West Greek (ancestors of the Dorians and of their relatives) were still living around the Pindus Mountains. In Eastern Thessaly, Boeotia, Attica, and the Peloponnesus and on certain Aegean islands (notably Crete), only varieties of Greek other than West Greek were spoken. The tablets reveal a somewhat artificial chancellery language current in the palace offices and taught as a written language in scribal training. Based essentially on a dialect of the type that was eventually called Arcado-Cypriot, it shows great uniformity in time (during the two centuries or thereabouts covered by documents) and in space (from Knossos to Thebes). The language of the tablets must have been based on the language of the Achaean rulers who inhabited the Mycenaean palaces. The problem of the geographic distribution of the various other forms of spoken Greek in Mycenaean times does not yet have a final solution.

Influence of Dorian invasions and Greek colonization

There followed two great events that upset the dialectal distribution within the Greek world. First, about 1100 BC the Dorian invasions brought speakers of West Greek southward, then into the Peloponnese, and finally into the Aegean. Some pre-Dorian Greek populations were expelled from their homes and emigrated eastward to the west coast of Anatolia and to Cyprus. Others, who remained where they were, became more or less thoroughly Dorian in speech. It has long been thought that some of the features that West Thessalian and, even more, Boeotian (both of which are Aeolic) shared in the 1st millennium with West Greek can be attributed to "recent" influences; on the other hand, some Doric dialects of the 1st millennium (*e.g.,* in Crete) show sporadic traces of features attributable to an Arcado-Cypriot substratum. The other subsequent event, which is of a different sort, was the great colonization movement that began in the 8th century BC. Each group of emigrants took the speech of its mother city and planted it in the new foundation. Thus there developed side by side on the shores of southern Italy and Sicily a totally new grouping of Greek dialects—Euboean Ionic at Cumae; Laconian Doric at Tarentum and Heraclea; Achaean at Sybaris, Croton, and Metapontum; Locrian at Locri Epizephyrii; Corinthian Doric at Syracuse; and so on.

Toward the middle of the 1st millennium BC the geographic distribution of the dialects (insofar as they are known directly through inscriptions) is, briefly, as follows:

West Group (Doric in the widest sense): (1) North-West Greek: Aetolia, Locris (colony—Locri Epizephyrii), Phocis, (2) Saronic Doric: the territory of Corinth (colonies—Corcyra, Syracuse), the Megarid (colonies—Megara Hyblaea, Selinus, Byzantium), Eastern Argolid, (3) Western Argolid, (4) South-East Aegean Doric: Melos and Thera (colony—Cyrene), Cos, Rhodes (colonies—Gela, Acragas), (5) Crete, (6) Laconia (colonies—Tarentum, Heraclea), Messenia, (7) Achaea (colonies—Ithaca, Sybaris), (8) Elis.

Aeolic Group: (1) Boeotia, (2) Thessaly, (3) Lesbos and Asiatic Aeolis.

Arcado-Cypriot Group: (1) Arcadia, (2) Cyprus, (3) Pamphylia (mixed with West Greek and Aeolic).

Ionic-Attic Group: (1) Attica, (2) Euboea (colonies—Catana, Cumae), (3) Northern Cyclades, (4) Asiatic Ionia (colonies—foundations in Pontus [Black Sea]).

This linguistic circumstance in the first half of the 1st millennium BC caused literature to develop on a dialect basis. The Homeric epic displays an artificial dialect based on Asiatic Ionic but interspersed with Aeolic and even Mycenaean elements inherited from a long oral tradition. Choral lyric also uses an artificial dialect, which is based on Doric but interspersed with many elements from Ionic epic and some from Lesbian poetry. Prose developed first in Ionic surroundings (Herodotus, Hippocrates), then in Attica (Thucydides, Plato). The dialect of dialogue in Attic tragedy is Attic mixed with some elements from choral lyric poetry. Attic comedy is pure Attic, whereas a Doric comedy developed in Syracuse. Personal poetry employs, depending on the author, Ionic (Archilochus, Hipponax), Lesbian (Alcaeus, Sappho), Boeotian (Corinna), and other dialects. It was only in the Hellenistic and Roman periods that Attic became clearly dominant, though in poetry of the later periods artificial imitations of the early genres were common.

Pre-Koine Ancient Greek

Within the alphabetical period of pre-Koine Ancient Greek (8th–4th centuries BC) there is no break between what is termed Archaic Greek (8th–6th centuries) and what is termed Classical Greek (5th–4th centuries).

In the linguistic subdivision of Ancient Greek the effects of substratum languages played only a minor part. In their penetration into Greece toward the beginning of the 2nd millennium BC, the Proto-Greeks found earlier populations established there, about whom Greek tradition gives only vague hints, and whose language or languages are unknown. From this "pre-Hellenic" stratum, Greek vocabulary made numerous borrowings (*kyparissos* 'cypress,' *asaminthos* 'bathing tub,' and so on), and it received from it a number of place names (*e.g.*, Korinthos); but there is no reason to think that the divergent characters of the Greek dialects (in phonetics and morphology) could be connected with different substrata. The native "barbarian" languages also had little effect on colonial Greek in the 1st millennium, and these contacts show up only in a few local borrowings.

On the other hand, there is a connection between the facts of civilization (in the political and cultural fields) and the evolution of the language. In the Mycenaean period an evident unity of civilization and the organization in the palaces of record offices and scribal training allowed the use of a stable and uniform chancellery language. In the first half of the 1st millennium, political subdivision and rivalry between cities allowed dialectal peculiarities to strike deep roots. As a result of a long oral formulaic tradition of epic poetry in dactylic hexameters, the language of the Homeric epics is an artificial mixture of dialects. This Homeric dialect became the standard for dactylic poetry all over the Greek world. In the 6th century BC, Ionic became the dialect of prose as a consequence of the political and cultural prestige of Ionia. In the course of the 5th century BC, as the prestige of Athens rose, Ionic was gradually replaced by Attic.

Linguistic characteristics. *Phonology.* The phonological systems of Ancient Greek differ noticeably from one period to another and from one dialect to another. The system that has been chosen to serve as an example here is that which may be attributed to Old Attic of about 500 BC.

In Old Attic there are seven vowel qualities: *i*, open and closed *e*, *a*, open and closed *o*, and *u*, each of which has a long and a short form, except open *e* and open *o*, which have only the long form. Diphthongs originally included *ei, ai, oi* and *eu, au, ou*, but *ei* began to evolve toward long closed *ē* and *ou* toward long closed *ō*. In addition, there is a diphthong *ui*, and, usually at the end of words, the diphthongs *-ēi, -āi, -ōi*, with long first elements, which much later were reduced respectively to long *ē*, long *ā*, and long open *ō*.

The consonantal structure is characterized by relative richness in stops (sounds produced by momentary complete closure at some point in the vocal tract): unvoiced *p, t, k*, aspirated *ph, th, ch*, voiced *b, d, g;* and by few spirants: only *s* and *h* (*h* restricted to initial position before a vowel). There are two liquid sounds, *l* and *r* and two nasals, *m* and *n*. The velar nasal (as in *ink*) is not distinctive but is only a variant of the *n* in front of a velar stop, or a variant of *g* in front of a nasal. Neither *y* nor *w* occurs as a distinctive sound. Most consonants can be doubled between vowels. The only consonant sounds normally allowed at the end of the word are *s, n*, and *r*.

Word accent

Apart from some unaccented monosyllabic or disyllabic terms of minor importance, each word is marked by an accent (the highest tone within the word) on one of the vowels (one of the last three vowels, if the word has more than three syllables). Short vowels, if they carry the accent, have only a rising tone (noted from the Alexandrian grammarians onward by [´] the sign of the acute accent); long vowels or diphthongs may have either a rising tone (noted by the acute accent) or a rising tone followed by a falling tone (noted by [ˆ] the circumflex). When a word carrying an acute accent on the vowel of the final syllable is followed by another word within the same phrase, its accent is noted by [`] the sign of the grave accent, in order to indicate that its tone is lower than that of the vowel of the initial syllable of the next word. Sometimes two otherwise identical words are differentiated by the nature or the position of the accent: for example, *oîkoi* ('houses') is a nominative plural form and *oíkoi* ('at home') an adverb of place; *tómos* means 'a cut' and *tomós* 'cutting.'

The accent (which is not associated with stress) does not play any part in the rhythm of the language. The rhythm of both prose and poetry is based upon the distribution of short and long syllables. For a syllable to be short, it must end in a short vowel; syllables ending in a long vowel, or closed syllables (*i.e.*, those ending in a consonant), are long. The rhythm of Ancient Greek is therefore said to be quantitative.

Morphology. Every nominal (noun or adjective) or verbal form combines a stem that carries the lexical sense of the word and a certain number of grammatical markers that serve to specify the meaning of the whole word (*e.g.*, plural, future) or to indicate its syntactic function (*e.g.*, subject, object) in the sentence.

The category of gender, which differentiates masculine, feminine, and neuter, concerns only the substantive (noun), adjective, and pronoun. The category of person (first, second, and third person) is restricted to the personal pronoun and the verb. There are three numbers—singular, dual, and plural—that are distinguished in both the noun and the verb. The survival of the dual is an archaism; although a living form in the Mycenaean period, it tends to be replaced by the plural in the 1st millennium. Attic is one of the dialects in which the dual number is best preserved down to the threshold of the Hellenistic period.

Not counting the vocative case, the Greek declension in the Mycenaean period still contained five cases: nominative, accusative, genitive, dative-locative, and instrumental. Between the Mycenaean period and the 8th century the instrumental ceased to exist as a distinct case, its role having been taken over by the dative.

The Greek verb system

The verb system is organized around four principal tense stems, which are built on the verb stem: "present," aorist, "perfect," and future. The first three are often called aspects, a term taken over from Slavic grammar. According to this terminology, the "present" stem is used for imperfective aspect (ongoing or repeated process), the "perfect" stem for stative aspect (state resulting from the completion of the process), and the aorist stem for perfective aspect (completed process). In principle, each tense stem provides five moods for the finite verb (with personal endings), a participle, and an infinitive. There are two assertive moods (nonpast and past indicative) and three nonassertive moods (subjunctive, optative, imperative). The expression of time relations is based on the combination of the values of the tenses ("aspects") and those of the moods. For instance, the past indicative of the "present" (the so-called imperfect) is used for ongoing or repeated processes in the past, that of the aorist for completed processes in the past, and that of the "perfect" (the so-called pluperfect) for states in the past. The Greek verb has two voices, active and mediopassive (reflexive and passive), which are expressed (leaving aside the aorist passive) by the opposition of two series of personal endings, for each mood (with participle and infinitive) of each tense stem.

Syntax. Since syntactical relations are expressed by means of case endings (and so on), Greek word order is relatively free. The creation of the definite article (post-Mycenaean and post-Homeric) is an important innovation. The availability of infinitive and participle clauses, with or without the article, as alternatives for all kinds of subordinate clauses permits the construction of very long and complex sentences that are nevertheless entirely transparent as to their syntactic structure. This accomplishment of Attic prose (known as periodic style) is unmatched in other languages. (M.Le./C.J.R.)

THE MIDDLE PHASES: KOINE AND BYZANTINE GREEK

Koine. The fairly uniform spoken Greek that gradually replaced the local dialects after the breakdown of old political barriers and the establishment of Alexander's empire in the 4th century BC is known as the Koine (*hē koinē dialektos* 'the common language'), or "Hellenistic Greek." Attic, by virtue of the undiminished cultural and commercial predominance of Athens, provided its basis; but, as the medium of communication throughout the new urban centres of Egypt, Syria, and Asia Minor, it absorbed numerous non-Attic elements and underwent some degree of grammatical simplification. Numerous inscriptions enable scholars to trace its triumphant progress at the expense of the old dialects, at least as the language of business and administration, although some rural dialects are reported to have survived as late as the 2nd century AD. Other sources of information for the Koine are the translation of the Septuagint made in the 3rd century BC for the use of the Hellenized Jewish community of Alexandria, the New Testament, and the writings of a few people (*e.g.*, the historian Polybius and the philosopher Epictetus) who preferred it to Attic. As the everyday colloquial language of urban Egypt, it may be studied in papyri going back to the 4th century BC. The Koine may be dated very crudely from the period of Alexander's conquests in the 4th century BC to approximately the reign of Justinian in the 6th century AD.

Changes from Ancient Greek to the Koine

The Koine replaced the Attic *tt* with the *ss* characteristic of Ionic and other dialects (*e.g.*, *glōssa* for *glōtta* 'tongue') at an early date, but its main phonological characteristic is the gradual simplification of the rich vowel system of Classical Greek. Ancient closed and open long /ē/ (*ει* and *η*) and /i/ (*ι*) merged as /i/, and /ai/ (*αι*) monophthongized to /e/; /oi/ (*οι*) monophthongized to /ü/, thus merging with simple /ü/ (*υ*) (pronounced as French *tu*). The second element of /au/ (*αυ*) and /eu/ (*ευ*) was changed to /v/ or /f/ depending on the voicing of the following consonant (compare Ancient *auge* 'light, dawn,' *autos* 'he' to Modern *avghí, aftós*). Classical /ph th kh/ (pronounced as in English *pin, tin, kin*) acquired fricative articulations as in *fin, thin,* and the final element of Scottish *loch* (or German *Buch*); /b d g/ became the voiced fricatives /v dh/ (as in *that*), /gh/ (as in Spanish *fuego*).

Other parts of the grammar also began to move in the direction of Modern Greek in this period. Nouns in consonant stems began to acquire the endings of the *-a* declension; *e.g., thygatēr,* 'daughter,' accusative *thygatera,* was remodeled after items such as *khōra, khōran* 'country.' The dual number was lost in nouns, verbs, and adjectives, as was the optative mood (expressing wish or desire) of verbs. Confusion arose between the perfect and aorist tense forms, leading to the loss of one or the other (the former in most verbs).

In vocabulary there were numerous borrowings from non-Attic dialects, and some Attic words acquired new meanings; thus *opsaria* 'fish' and *brechei* 'it rains' for Classical Greek *ichthyes* and *hyei* both occur in the New Testament (compare Modern Greek *psárya, vrékhi*).

This gradual divergence from the language of Plato and Demosthenes was viewed as a species of linguistic decadence by an influential school known as the Atticists, who unceasingly castigated the use of Koine forms by writers. It was thus that the rift developed between the everyday spoken language and an archaizing, specifically written language. It became fashionable to publish manuals of "good usage" in which the Attic equivalents of Koine innovations were recommended as models for the student's imitation.

Byzantine Greek. During the period of the Byzantine Empire (*i.e.,* until the fall of Constantinople in 1453) the language of administration and of most writing was firmly rooted in the Atticist tradition; it is this archaizing style that is often referred to as "Byzantine Greek." The spoken language continued to develop apace, however, and its course can be followed to some extent in the writings of the less-educated chroniclers (such as John Malalas, 6th century) and hagiographers. Furthermore, the increasing political and military disintegration characteristic of the last few centuries preceding the fall of Constantinople brought with it a general decline in educational level, and works appeared that reflect quite closely the colloquial language of the time, although learned and pseudo-learned elements are never absent. While the differences between the *Chronicle of the Morea* (13th century), for example, and present-day spoken Greek are quite minor, Byzantium failed to produce a writer of the stature of Dante, capable of establishing once and for all the living vernacular as a worthy vehicle for great literature.

Changes occurring in Byzantine Greek

Most of the phonological and grammatical developments that separate present-day Greek from the Koine occurred during this period. Thus, in the phonology the two high front vowels /i/ and /ü/ were merged, simplifying the six-vowel system to the five-vowel system of Modern Greek. In the morphology the frequent misuse of the dative case of nouns shows that it went out of use in the spoken language, and the infinitive was replaced by various periphrastic constructions. (Periphrastic constructions involve the use of function words and auxiliaries.) In the earlier period numerous words (mostly Latin) were imported: the chronicler Malalas has (in their modern form) *pórta* 'door,' *kâmbos* 'plain,' *saíta* 'arrow,' *paláti* 'palace,' *spíti* 'house' (from *hospitium*), and hundreds of other borrowings, not all of which have survived. The later period is characterized by the richness of its compound words, usually from native roots. Some of these, such as the compounds in which a modifying noun precedes its head noun, continued ancient patterns (*thalassóvrakhi* 'sea rock,' *vunópulo* 'mountain lad'); coordinative compounds of the type common in Modern Greek, though rare in earlier periods, are also found (*aristódhipnon* 'lunch and dinner,' compare Modern Greek *andróyino* 'man and wife,' *makheropíruna* 'knives and forks'). Semantic shift was another source of innovation: *álogho* 'horse,' previously meant 'irrational'; *skiâzome* 'I fear,' earlier meant 'I am in shadow'; and (*u*)*dhén* 'not,' meant, in Classical Greek, 'nothing.'

MODERN GREEK

History and development. Modern Greek derives from the Koine via the local varieties that presumably arose during the Byzantine period and is the mother tongue of the inhabitants of Greece and of the Greek population of the island of Cyprus. Before the population exchange in 1923, there were Greek-speaking communities in Turkey (Pontus and Cappadocia). Greek is also the language of the Greek communities outside Greece, as in the United States, Canada, and Australia. There are Greek-speaking enclaves in Calabria (southern Italy) and in Ukraine. Two main varieties of the language may be distinguished: the local dialects, which may differ from one another considerably, and the Standard Modern Greek (Greek: *Koini Neoelliniki,* "Common Modern Greek").

Local dialects. Of the local dialects, Tsakonian, spoken in certain mountain villages in eastern Peloponnese, is quite aberrant and shows evidence of descent from the ancient Doric dialect (*e.g.,* it often has an /a/ sound for the early Greek /ā/ that went to /ē/ in Attic, later to /i/). The Asia Minor dialects also display archaic features (*e.g.,* Pontic /e/ for ancient /ē/ in certain words). It is not certain whether southern Italian Greek represents a survival from ancient times or was reimported there during the Byzantine period. Apart from these peripheral varieties, the modern dialects may be grouped for practical purposes as follows:

Divisions of the modern dialects

1. Peloponnesian, differing but slightly from the dialects of the Ionian isles, forms the basis of standard Demotic. It shows very few specifically local innovations in its phonology, although its verb morphology is less conservative than that of the island dialects.

2. Northern dialects, spoken on the mainland north of Attica, in northern Euboea, and on the islands of the northern Aegean, are characterized by their loss of unstressed /i/ and /u/ and the raising of unstressed /e/ and /o/ sounds to /i/ and /u/. Thus, standard *kotópulo* 'chicken' becomes *kutóplu, émine* 'he stayed' becomes *émni.* They also mark certain first and second person plural past tense verb forms with *-an* (*ímastan* 'we were,' Athenian *ímaste*) and use the accusative for indirect object pronouns where the southern dialects have the genitive (*na se pô* 'let me tell you,' standard *na su pô*).

3. Old Athenian was spoken in Athens itself until 1833, when Athens became the capital of the modern state, and on Aegina until early in the 20th century; a few elderly speakers still remain in Megara and in the Kími district of central Euboea. Its salient feature is the replacement of the Byzantine /ü/ sound (from ancient /ü/, /oi/) by /u/

rather than normal /i/; it changes the /k/ sound before the vowels /e/ and /i/ to /ts/ and fails to contract the vowels /i/ and /e/ to a /y/ sound before vowels (ancient *sykéa* becomes *sutséa* 'fig tree,' standard *sikyá*).

4. Cretan softens /k/ to a /č/ sound (as in *church*), /kh/ to /š/ (as in *she*) before /i/ and /e/, and /y/ to /ž/ (as the *s* in *pleasure*)—*e.g., če* 'and,' *šéri* 'hand,' *žéros* 'old man,' standard *ke, khéri, yéros*.

5. The southeastern dialects of Cyprus, Rhodes, Chios, and other islands in the area also soften /k/ to /č/, drop voiced fricative consonants between vowels, and retain the ancient final *-n* (*láin* 'oil,' standard *ládhi*). They also retain the contrast between long and short consonants (*fíla* 'kiss [imperative]' but *fílla* 'leaves'). As is done in Cretan and Old Athenian, they add *gh* to the suffix *-ev-* that occurs at the end of many verb stems (*dhulévgho* 'I work,' standard *dhulévo*).

Standard Modern Greek. With the establishment of the new Greek state in 1830, the Peloponnesian dialect was adopted as the oral language, and this developed into the Demotic variety. The absence of a written form of Demotic, however, led to the creation of Katharevusa, a "pure," rather artificially archaizing form that was intended to purge the language of foreign elements and to systematize its morphology (inevitably on the Classical Greek model). This Modern Greek "diglossia" continued well into the 20th century, with specific areas of use for the two varieties; *e.g.,* Demotic became the vehicle for poetry, whereas Katharevusa remained the language of administration.

The emergence of Standard Modern Greek

The diglossia problem was finally resolved in 1976, when Demotic was declared the official language of Greece. Meanwhile, the two varieties had naturally converged, and the emerging Standard Modern Greek language can be well characterized as resulting from the merger of the Demotic variety with Katharevusa features. Thus, in the phonology, some clearly Demotic changes (see statements [3] and [5] below), under Katharevusa influence, either were suppressed or developed alternations or even contrasts. Thus, for the change of *i* to *y* before another vowel, is found *jimnásio* 'high school' (instead of the expected Demotic **jimnásyo*), the alternants *sxolyó* and *sxolío* 'school,' or the contrast *yós* 'sun' but *iós* 'virus.' The assimilation of a nasal to a fricative is confined to the morphology; *e.g.,* the verbal form *krin-thik-e* (third person-passive-singular aorist 'he was judged') is *kríthike* with nasal loss, while in a word such as *pénthos* 'mourning' the nasal is retained. Further, the historical tendency to differentiate gender by declension class (*e.g.,* by restricting the declension *-os* to masculine) was inhibited, and numerous feminine nouns in *-os* are reintroduced into the language (*e.g., odhós* 'street,' *leofóros* 'avenue'), some in parallel with Demotic alternants, as in *jatrós* or *jatrína* 'doctor (feminine).'

The interaction between Demotic and Katharevusa is even stronger in the vocabulary. A Katharevusa form may be used in parallel with a Demotic form in a specialized role; for example, *édhra* (from the Ancient Greek word for 'chair') means 'professorial chair,' while the Demotic *karékla* remains in use for the article of furniture.

This manner of interaction causes the Greek speaker to experience these differences not as belonging to two different varieties of the language but rather as stylistic variations of one and the same system.

Linguistic characteristics. *Phonology.* Modern Greek has five distinct vowel sounds (/i/, /e/, /a/, /o/, /u/) and the glide /y/, most of which are indicated in Greek orthography in more than one way. The consonant sounds are:

Voiceless stops	p	t		k
Voiced stops	b	d		g
Voiceless fricatives	f	th	s	kh
Voiced fricatives	v	dh	z	gh
Nasals	m	n		
Liquids		l, r		

Historically, /f th kh/ derive from ancient aspirated consonants, and the voiced fricatives /v dh gh/ from voiced stops /b d g/. Modern /b d g/ usually result from the voicing of /p t k/ after nasals; thus Ancient Greek *pente* 'five' becomes *pénde.* They also occur at the beginning of words in place of ancient nasal + stop sequences (*boró* 'I am able' from *emporó*). Other important consonant cluster changes linking Ancient and Modern Greek include:

1. Ancient clusters, whether of stops or of aspirates, become fricative + stop; for example, *hepta* 'seven' becomes *eftá, (e)khthes* 'yesterday' becomes *(e)khtés.*

2. Double consonants are simplified except in the southeast, thus *thalassa* 'sea' becomes *thálasa.*

3. Nasals assimilate to the following fricatives; thus *nymphē* 'bride' becomes *níffi* and then (except in the southeast dialects) *nífi.*

4. The liquid /l/ may be replaced with /r/ before consonants; for example, *adelphos* 'brother' becomes *adherfós.*

5. Before a vowel, /i/ and /e/ change to /y/; thus *paidia* 'boys' becomes *pedhyá. mēlea* 'apple tree' becomes *milyá.* Except for the simplification of double consonants, these historical changes do not hold for words of Katharevusa origin.

Predictable stress placement

With the changes produced in the vocalic system in Koine, the ancient pitch distinction was lost and stress became dynamic (as in English), its place being indicated orthographically by a uniform stress mark; but it remained confined to the three last syllables of a word (the trisyllabic, or window, constraint). Stress placement is largely predictable, depending for nominals on their declensional class marker (*e.g., ánthropos* 'man' versus *polítis* 'citizen' [*-o* versus *-i* class]), but for the verb on their tense (*e.g., katháriz-a* 'I cleaned' versus *katharíz-o* 'I clean' [past versus nonpast tense]).

Further stress shift may occur owing to the trisyllabic constraint, as in *máthima* gives *mathímata* 'lesson' (nominative singular or plural), or as a morphological relic of an earlier long ō-vowel in the genitive plural—*e.g., mathímata* becomes *mathimáton* 'lesson' (nominative or genitive plural). The addition of clitics (words that are treated in pronunciation as forming a part of a neighbouring word and that are often unaccented or contracted) may provoke further stressing in the host + clitic unit if the trisyllabic constraint is violated, as in *máthima*: but *máthima-mu* becomes *máthimá mu* 'lesson' becomes 'my lesson.' In some dialects, especially in the north, the tendency to a rhizotonic (stable) stressing extends to the verb, leading either to violations of the trisyllabic constraint or to an additional stress (as in the case of clitics)—*e.g., tarázumasti* or *tarázumásti* 'we are shaken' (standard *tarazómaste*).

Morphology and syntax. Much of the inflectional apparatus of the ancient language is retained in Modern Greek. Nouns may be singular or plural—the dual is lost—and all dialects distinguish a nominative (subject) case and accusative (object) case. A noun modifying a second noun is expressed by the genitive case except in the north, where a prepositional phrase is usually preferred. The indirect object is also expressed by the genitive case (or by the preposition *se* 'to,' which governs the accusative, as do all prepositions). Thus:

o yatrós	*ípe tin istoría*	*s ton adherfó tiz dhaskálas*
'The doctor	told the story	to the brother of the teacher'
(nominative)	(accusative)	(genitive)

The ancient categorization of nouns into masculine, feminine, and neuter survives intact, and adjectives agree in gender, number, and case with their nouns, as do the articles (*o* 'the,' *enas* 'a'). In general, pronouns exhibit the same categories as nouns, but the relative pronoun *pu* is invariant, its relation to its own clause being expressed when necessary by a personal pronoun in the appropriate case: *i yinéka pu tin ídhe to korítsi* 'the woman *pu* her saw the girl' (*i.e.,* 'the woman whom the girl saw').

The verb is inflected for mood (indicative, subjunctive, imperative), aspect (perfective, imperfective), voice (active, passive), tense (present, past), and person (first, second, and third, singular and plural). The future is expressed by a particle *tha* (from earlier *thé[o] na* '[I] want to') followed by a finite verb—*e.g., tha grápho* 'I will write.' Formally, the finite forms of the verb (those with personal endings) consist of a stem + (optionally) the perfective aspect marker (*-s-* in active, *-th-* in passive) + personal ending (indicating person, tense, mood, voice). Past forms are prefixed by *e-* (the "augment"), usually lost in mainland dialects when unstressed. There are also two nonfinite forms, an indeclinable present active participle in *-ondas* (*ghrâfondas* 'writing'), and a past passive one in *-mênos* (*kurazmênos* 'tired').

Aspect in Modern Greek

Aspectual differences play a crucial role. Roughly, the perfective marker indicates completed, momentary action; its absence signifies an action viewed as incomplete, continuous, or repeated. Thus the imperfective imperative *ghrâphe* might mean 'start writing!' or 'write regularly!' while *ghrâpse* means rather 'write down! (on a particular occasion).' Compare also *tha ghrâpho* 'I'll be writing' but *tha ghrâpso* 'I'll write (once).' The difference is sometimes represented lexically in English: *ákuye* 'he listened' and *ákuse* 'he heard.' The passive forms are largely confined to certain verbs active in meaning like *érkhome* 'I come,' *fovâme* 'I am afraid,' and reciprocal usages (*filyóndusan* 'they were kissing').

The most common form of derivation is by suffixation; derivation by prefixation is limited mainly to verbs. On the other hand, compound formation is rich. Three morphological types of compounds can be distinguished, as reflected also in their stressing—thus, stem + stem compounds—*e.g., palyófilos* 'old friend' (*o* is the compound vowel) or *khortofághos* 'vegetarian'; stem + word compounds—*e.g., palyofílos* 'lousy friend' (compare *filos* 'friend'); and the newly borrowed formation, word + word compounds—*e.g., pedhí thâvma* as English 'boy wonder.' There is no infinitive; ancient constructions involving it are usually replaced by *na* (from ancient *hína* 'so that') + subjunctive. Thus *thélo na ghrâpso* 'I want to write,' *borí na ghrâpsi* 'he can write.' Subordinate statement is introduced by *óti* or *pos* (*léi óti févghi* 'he says that he is leaving'). Unlike English, Greek (because of its inflectional system) shows flexible word order even in the simplest sentences. Further, the subject of a sentence may be omitted.

Vocabulary. The vast majority of Demotic words are inherited from Ancient Greek, although quite often with changed meaning—*e.g., filó* 'I kiss' (originally 'love'), *trógho* 'I eat' (from 'nibble'), *kóri* 'daughter' (from 'girl'). Many others represent unattested combinations of ancient roots and affixes; others enter Demotic via Katharevusa: *musío* 'museum,' *stikhío* 'element' (but inherited *stikhyó* 'ghost'), *ekteló* 'I execute.' In addition, there are more than 2,000 words in common use drawn from Italian and Turkish (accounting for about a third each), and from Latin, French, and, increasingly, English. The Latin, Italian, and Turkish elements (mostly nouns) acquire Greek inflections (from Italian *síghuros* 'sure,' *servitóros* 'servant,' from Turkish *zóri* 'force,' *khasâpis* 'butcher'), while more recent loans from French and English remain unintegrated (*spor* 'sport,' *bar* 'bar,' *asansér* 'elevator,' *futból* 'football,' *kompyúter* 'computer,' *tênis* 'tennis').

(B.E.N./A.M.-D.)

Italic languages

Italic languages, in a broad sense, are certain Indo-European languages that were once spoken in the Apennine Peninsula (modern Italy) and in the eastern part of the Po valley. These include the Latin, Faliscan, Osco-Umbrian, South Picene, and Venetic languages, which have in common a considerable number of features that separate them from the other languages of the same area—*e.g.,* from Greek and Etruscan. (In a more narrow sense, the term Italic languages excludes Latin and denotes only Osco-Umbrian, South Picene, Faliscan, and Venetic.)

For a long time the Italic languages have been considered to be an Indo-European subfamily like Celtic, Germanic, or Slavic. Today some scholars are inclined to distinguish within the so-called Italic branch at least three independent members of the Indo-European family: Latin (with Faliscan), Osco-Umbrian (with South Picene), and Venetic (if indeed this is an Italic language, as will be assumed in this article). They attribute the similarities—*i.e.,* the unifying phenomena in the division—to a convergence that took place when the speakers of these different idioms were integrated into the "Italic" civilization of the early first millennium BC. The culture that resulted is known as the "Etruscan koine." Figure 4 shows the assumed distribution of languages in ancient Italy.

Figure 4: Supposed language areas of the Italic and neighbouring languages about 250 BC.

DISTRIBUTION AND ORIGINS

Languages of the group. *Latin.* Latin is the language of Latium and of Rome; its earliest known documents date to the 6th century BC. Rich epigraphic evidence and an extensive literature began at the end of the 3rd century BC, at the time when Roman Latin was emerging as the predominant language of Italy. By AD 100 at the latest, Latin had effaced all the other dialects between Sicily and the Alps, with the exception of Greek in the colonies of Magna Graecia. (For more information about Latin and about the languages that derive from it, see below *Romance languages.*)

Spread of Latin on the Italian peninsula

The other Italic languages—Italic languages in the narrow sense—are known through local and personal names transmitted by Greek and Roman sources, and especially from inscriptions.

Oscan. Before Latin spread out, Oscan was the most widely spoken group of dialects of the Apennine Peninsula. It was used by the Samnites in Samnium and Campania; by the inhabitants of Lucania and Bruttium; and, with slight variations, by smaller tribes between Latium and the Adriatic coast: the Volsci, Marsi, Paeligni, Vestini, and Marrucini. (These distinct "minor dialects" within the Oscan language continuum are referred to as Volscian, Marsian, Paelignian, etc., and collectively as Sabellic, or Sabellian, dialects; this term can also refer more generally to the Italic languages other than Latin and Venetic, roughly comparable to the narrow sense of the term Italic languages.) The legendary Sabines, who shared the earliest history of Rome, probably also spoke an Oscan dialect. The most important Oscan texts come from Campanian cities, as do the oldest ones (several so-called Paleo-Oscan, or "Pre-Samnite," vessel inscriptions of the form "I belong to So-and-so," dating to as early as the mid-6th century BC). The largest text, a treaty between Nola and Abella from the 2nd century BC, is carved on a stone slab, called the Cippus Abellanus. In Bantia, a nearly unknown town of Lucania, the Tabula Bantina is preserved, the most extensive Oscan inscription. It is a bronze tablet with penal laws concerning municipal administration, written in Latin letters during the first half of the 1st century BC. The oldest Oscan text of any length is the so-called Agnone Tablet of about 250 BC (a small bronze tablet found near Fonte Romito, between Agnone and Capracotta), detailing cultic instructions related to the worship of Ceres and other divinities. The remainder of the Oscan corpus includes diverse material, some of which is of considerable cultural interest (short cultic and sacrificial texts, curse tablets, and several types of municipal inscriptions).

Umbrian. The Umbrian idiom, closely related to Oscan, is known from a few small inscriptions and from the Tabulae Iguvinae (Iguvine Tables), which consist of seven bronze tablets found at Gubbio (the ancient Iguvium). Constituting one of the largest and most important epigraphic documents of antiquity, the tablets contain ritual regulations of a sacred brotherhood to which a considerable part of the public cults of Iguvium was delegated. The Tabulae Iguvinae were incised, partly in the Umbrian alphabet and partly in Latin letters, within the last three centuries BC, but portions of the text itself result from a far more remote oral tradition.

Inscriptions in South Picene

South Picene. Approximately two dozen short inscriptions from the southern part of ancient Picenum (most of them found near Ascoli Piceno and Teramo) preserve extremely early remains (6th and 5th centuries BC) of the Italic language now known as South Picene. Formerly referred to as Old Sabellian or Central Adriatic, South Picene shows close affinities to Oscan and Umbrian. (North Picene, known mainly from a 12-line inscription of about 500 BC on a stele from Novilara, is not an Italic language and may not even be an Indo-European language.)

Faliscan. Faliscan inscriptions appear only in the immediate surroundings of Falerii (the present Civita Castellana in central Italy), which, except for its dialect, seems to have been a completely Etruscan city. The Faliscan corpus consists mainly of short sepulchral inscriptions, most of which date to the 4th century BC and later. There are, however, several important Archaic Faliscan vessel inscriptions (see Figure 7), including some from as early as the 6th century BC that show evidence of poetic composition.

Venetic. The language represented by inscriptions from the territory of the Veneti—between the Po River, the Carnic Alps, and Istria—is called Venetic. The majority of discoveries come from sanctuaries at Este and Làgole di Calalzo. The Venetic inscriptions (of which there are about 300, ranging from the 5th to the 1st century BC) consist almost entirely of brief epitaphs and dedications (compare the votive text in Figure 8).

Several very early texts cannot be ascribed with certainty to any of the individual languages listed above but show linguistic features of a broadly Osco-Umbrian/South Picene type. The most important of these is found on an inscribed flask from Poggio Sommavilla (in Sabine territory), dating from the end of the 7th century BC.

The scripts used for writing these languages include the Greek alphabet in Bruttium and Lucania and the Latin alphabet and various derivations of the Etruscan alphabet in the other regions. Five "national" or "native" alphabets are distinguished: Oscan, Umbrian, South Picene, Faliscan, and Venetic (see Figures 5–8 for four representatives).

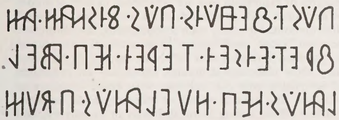


Figure 5: *Oscan.*
Inscription from the Cippus Abellanus: *púst.feihúís.pús. físnam.am/fret.eiseí.tereí.nep.abel/lanús.nep.núvlanús. pídum/ . . . [ú represents o].* (Latin *Post muros qui fanum circumdant, in eo territorio neque Abellani neque Nolani quicquam [aedificaverint].*) "Behind the walls which go around the sanctuary,—in this area neither the inhabitants of Abella nor the inhabitants of Nola [are allowed to construct] anything."

Origins of the Italic languages. The Italic languages must have been brought from the original area of the Indo-European languages, perhaps in eastern parts of central Europe, when their speakers crossed the Alps. This is attested to by a stratum of very old place-names of non-Indo-European origin—*e.g.,* Tarracina, Capua—that covers not only the Apennine Peninsula but also Greece and Anatolia. This stratum is ascribed to a "Mediterranean" language believed to have dominated large parts of the ancient world before the arrival of the Indo-European peoples. Nothing is known about the date, the path, and the circumstances of the above-mentioned immigration, and none of the many attempts to combine archaeological evidence with linguistic prehistory has led to convincing results. Thus, the only resources available for studying the Italic languages are exclusively linguistic methods of comparative philology.

Possible "Mediterranean" stratum

LINGUISTIC CHARACTERISTICS

Phonology. Many of the phonetic processes that make the attested Italic languages differ from the reconstructed Indo-European language seem to have occurred relatively late in time. The only one that can confidently be placed outside of Italy—that is, before the immigration over the Alps—is the change to *ss* in combinations of *d* (dental occlusive, or dental stop) + *t*. This is a feature common to Celtic, Germanic, and the Italic languages. For example, Latin *visus* comes from the older, reconstructed form **wissos* 'seen'; this is cognate with High German *gi-wiss* 'surely known' and Old Irish *ro-fess* 'is known,' all of these forms deriving from an Indo-European term **wid-to-s,* with $d + t$. (An asterisk [*] before a word means that it is not attested but reconstructed.)

The development of the Indo-European labiovelar stop k^w is more complex. (A labiovelar stop is a sound pronounced with simultaneous articulation—movement—of the lips and the velum, the soft palate.) From this sound there has resulted a *qu* in Latin, *p* in Osco-Umbrian and South Picene, *c* in Irish, and *p* in Brythonic Celtic; *e.g.,* Latin *quis* 'who(ever)' is cognate with Oscan *pis* and Umbrian *pis* (similarly South Picene *pim* 'whom[ever]' or 'which[ever]'), these forms deriving from Indo-European **$k^w$is;* and Irish *cia* is related to Welsh *pwy,* 'who,' derived from Indo-European **k^wei.* Some scholars have tried to trace this development back to an Italo-Celtic unity, but the change of Brythonic k^w to *p* is surely later than the dropping of the *p* in Common Celtic. It is sounder, therefore, to assume independent processes in the different languages.

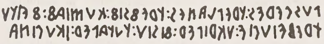


Figure 6: *Umbrian.*
Passus from the Iguvine Tables: *pus veres treplanes tref sif kumiaf feitu/trebe iuvie ukriper fisiu tutaper ikuvina [u represents u and o, k represents k and g].* (Latin *Post portam Trebulanam tres sues gravidas facito Trebo Iovio pro arce Fisia, pro civitate Iguvina.*) "Behind the Trebulan gate he shall sacrifice three pregnant sows to [the god] Trebus Iovius, for the Fisian citadel, for the state of Iguvium."

Other features developed in Italy itself—*e.g.,* the use of the voiceless dental spirant (fricative) *f* that is shared with Etruscan and is lacking in marginal districts of Venetic. In all Italic languages this *f* sound replaced the Indo-European voiced aspirated sounds (represented as *bh, dh, g^wh*) in initial position. Examples of the use of *f* in Italic are as follows: Latin *frater* 'brother' = Umbrian *frater,* from Indo-European **bhrātēr;* Latin *facio* 'I do, make' is related to Oscan *fakiiad* 'he should do' and Venetic *fagsto* 'he made,' based on an Indo-European stem **dhə-k-.* A more recent common process in Latin and Osco-Umbrian is the use of the full system of five short vowels in initial syllables only; short vowels of noninitial syllables in Latin became less open—*e.g., facio* 'I do, make,' but *inficio,* the compound of *in* + *facio.* In Osco-Umbrian these vowels tend to be lost completely—*e.g.,* Umbrian *benust* 'he will have come,' but Oscan *cebnust* 'he will have come near.' Some differences between Latin and Osco-Umbrian/South Picene probably arose during the last centuries BC—

The *f* sound

e.g., Osco-Umbrian/South Picene *ō* changed to *u* (Oscan *dúnúm,* South Picene *dúnoí,* Latin *dōnum* 'gift'), *ē* became *i* (Oscan *ligud,* Latin *lēge* 'law' in the ablative singular; South Picene *spolítiú,* Latin *Spolētium* [name of a town in Umbria, modern Spoleto]), and final *ā* developed into *o* (*viú* [*ú* in the Oscan national alphabet = *o*], Latin *via* 'way'). Indo-European voiced aspirated sounds (*bh, dh, gʷh*) in internal position probably first became voiced spirants (*e.g.,* sounds such as *v*) in all Italic languages and, later, voiced stops in Latin and Venetic and the voiceless spirant *f* in Osco-Umbrian, South Picene, and Faliscan. Examples of these changes (for Indo-European *dh*) are the voiced stop *b* in Latin *liberi* '(free) children' and Venetic *louderobos* 'children' (in the dative plural) versus the voiceless spirant *f* in Oscan *loufro-* 'free' and Faliscan *loferta* 'freed woman.' Examples for Indo-European *bh* are Oscan *tfei,* Umbrian *tefe,* South Picene *tefeí* versus Latin *tibi* 'to/for you.'

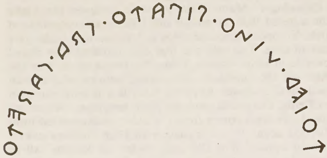


Figure 7: *Faliscan.*
Inscription on a bowl: *foied.uino.pipato.cra.carefo.* (Latin *Hodie vinum bibam, cras carebo.*) "Today I shall drink wine, tomorrow I shall have nothing."

Morphology. In contrast to the phonology, which shows so many correlations among the Italic languages, there are few definite connections between these tongues in their grammars. A characteristic innovation is the extension of the ablative singular case from *o*-stems and pronouns, where it occurred originally, to other declension classes: Latin *praidad* 'with the plunder,' later *praeda, meretod* 'by merit,' Oscan *toutad* 'by the people,' *slaagid* 'of the border,' South Picene *arítih* (*-h* represents earlier *-d*) 'with skill.' Many of the morphological features common to Osco-Umbrian, South Picene, and Latin are shared by other Indo-European languages; that is, they are not Italic in a specific sense. For example, the *a*-subjunctive—*e.g.,* Latin *faciat* 'may he do' and Oscan *fakiiad*—is also Celtic; passive endings in *-r*—*e.g.,* Oscan *vincter* and Latin *vincitur* 'he is conquered,' South Picene *qolofitúr* 'he honours/supports' or 'honour/support is shown to'—are found in Celtic, Hittite, and Tocharian as well. More important are the discrepancies. For example, the genitive singular of *o*-stems shows *-ī* in Latin, Faliscan, and Venetic and in the Celtic languages, but *-eis* in Osco-Umbrian and South Picene; the nominative plural of the same class is marked by *-oi* in early Latin, Celtic, and Greek but by *-ōs* in Osco-Umbrian, South Picene, Germanic, Sanskrit, and other languages. In addition, the perfect stems of secondary verbs (verbs derived from nouns or from other verbs) are formed by *-u-* or *-v-* in Latin, by *-t(t)-* in Oscan, by *-s-* in Venetic, and by a variety of formants in Umbrian—*e.g.,* Latin *donavit* 'he has given' = Oscan *duunated* = Venetic *donasto,* Umbrian *combifiansiust* 'he will have announced' (with formant *-nky-*).

Shared features

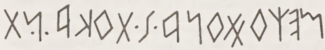


Figure 8: *Venetic.*
Inscription on a capital serving as pedestal of a votive statue found at Este: *mego donasto kanta.* (Latin *Me donavit Canta.*) "Canta gave me" ("to the goddess" is understood).
From G.B. Pellegrini and A.L. Prosdocimi, *La Lingua Venetica* (1967)

Vocabulary. Lexical comparison leads to more specific data about the history of the Italic languages. There are linguistic boundaries called isoglosses that may date back to pre-Italic history: *e.g.,* Oscan *humuns,* Latin *homines,* and Gothic *gumans* 'human beings' derive from an Indo-European root that meant 'earth'; and Oscan *anamúm* 'mind' (accusative singular) is directly related to Latin *animus* 'mind, soul' and Irish *anam* 'soul,' these words deriving from an Indo-European root meaning 'to breathe' (compare Greek *anemos* 'wind'). There are many old differences between Latin on the one hand and Osco-Umbrian (and South Picene) on the other. Latin *ignis* 'fire' = Sanskrit *agni,* but Umbrian *pir* 'fire' = Greek *pȳr* = Old English *fyr;* Latin *aqua* 'water' = Gothic *ahwa,* but Umbrian *utur* 'water' = Greek *hydōr* = Old English *wæter;* Latin *filius, filia* 'son, daughter,' but Oscan *puklu* and South Picene *puqloh* 'son' = Sanskrit *putra,* and Oscan *futír* 'daughter' = Greek *thygatēr* = Gothic *dauhtar.*

Certain lexical fields that reflect the acquisition of the Mediterranean culture show an independent terminology. The following forms strongly suggest that Latin and Osco-Umbrian speakers were not in contact with each other when they began to build cities: Latin *porta* 'gate,' Oscan *veru* 'gate'; Latin *arx* 'citadel,' Umbrian *ocar* 'citadel, castle'; Latin *moenia* 'walls, ramparts,' *murus* 'wall,' Oscan *feíhúss* (accusative plural) 'walls.' On the other hand, Latins and Osco-Umbrians adopted the same terms for 'write' and 'read'; Latin *scribere* 'to write,' Oscan *scriftas* 'written'; Latin *legere* 'to read,' Paelignian *lexe* 'you will read.' It is known that the Latin and Osco-Umbrian alphabets are derived from the Etruscan alphabet; the spread of these terms can, therefore, be attributed to a period of Etruscan predominance. Etruscan features are obvious in archaic Italic religion; Osco-Umbrians and Veneti adopted even the Etruscan word for 'god'—*ais.* Many other religious terms show a close community among Italic peoples; *e.g.,* the Latin forms *pius* 'pious, obedient' and *piare* 'to honour with religious rites' are equivalent to Volscian *pihom* (neuter singular) and Umbrian *pihatu* (imperative). These correspondences in religious vocabulary extend to whole phrases, some inherited from Indo-European forms of expression and therefore of very great antiquity: *e.g.,* the Umbrian prayer formula *ueiro pequo . . . salua seritu* 'keep safe men (and) livestock' is directly comparable to Latin *pastores pecuaque salva servassis* 'may you keep safe shepherds and livestock' (in an ancient Roman prayer), and both are cognate with similar expressions in Vedic Sanskrit and Avestan.

Etruscan influence

The Etruscan supremacy ended with the foundation of local republics in Rome and in other cities of Italy in approximately 500 BC; when that occurred, the unifying force of Etruscan culture lost its influence. Early republican terminology developed independently; *e.g.,* Latin *consul* 'consul,' but Oscan *meddíss* designate the first magistrate; to Latin *senatus* 'senate' corresponds Oscan *kúmparakineís* (genitive singular), and to Latin *comitia* 'assembly,' the Oscan forms *comono* or *kúmbennieís.* The last period of Italic language history is characterized by an increasing influence of Roman models. For example, Oscan *ceus* 'citizen' is a Latin loanword that stems from a form **ceuis,* which existed about 200 BC and was intermediate between Old Latin *ceivis* and its later form *civis;* Oscan *aídil* and *kvaísstur* imitate Latin *aedilis* and *quaestor,* terms for offices in the Roman government; and the Veneti adopted the Roman word for 'freed man,' *libertus.* In addition, the Oscan Tabula Bantina slavishly copied the juridical style and terminology of the Romans.

Poetics. Beyond vocabulary, the Italic languages share features of expression that are prominent in formal modes of discourse, such as prayers, oaths and curses, and legal formulations. For many such compositions, the distinction between "prose" and "poetry" is an artificial one: prayers or juridical formulas that otherwise appear to be prose may show elaborate structures dependent on sequences of isosyllabic cola (*i.e.,* phrases with equal numbers of syllables, like lines of verse) or other rhythmic patterns, marked effects arising from antithesis or synonymy, and a high degree of asyndeton (*i.e.,* lack of connective words such as *and*). Perhaps most striking is the phonetic feature of alliteration, pervasive throughout early Latin poetry, as well as in formal documents composed in Oscan, Umbrian, and South Picene. For Latin and Umbrian, compare the parallel phrases *salua seritu* and *salva servassis* in the prayer mentioned above, with the latter preceded by *pastores pecuaque;* an Oscan curse includes the sequence *fakinss, fangvam, biass, biítam* '(I hand over to the divinity [*i.e.*, I curse] his) deeds, tongue, strength, life'; and a South Picene epitaph contains the concluding sequence *yeiat yepetí* 'he lies in the tomb.' (J.U./B.Vi.)

Romance languages

The Romance languages, all derived from Latin within historical times, form a subgroup of the Italic branch of the Indo-European language family (see above *Italic languages*). The major languages of the family include French, Italian, Spanish, Portuguese, and Romanian, all national languages. Catalan also has taken on a political and cultural significance; among the Romance languages that now have less political or literary significance or both are the Occitan and Rhaetian dialects, Sardinian, and Dalmatian (extinct), among others. Of all the so-called families of languages, the Romance group is perhaps the simplest to identify and the easiest to account for historically. Not only do Romance languages share a good proportion of basic vocabulary—still recognizably the same in spite of some phonological changes—and a number of similar grammatical forms, but they can be traced back, with but few breaks in continuity, to the language of the Roman Empire. So close is the similarity of each of the Romance languages to Latin as currently known from a rich literature and continuous religious and scholarly tradition that no one doubts the relationship. For the nonspecialist, the testimony of history is even more convincing than the linguistic evidence; Roman occupation of Italy, the Iberian Peninsula, Gaul, and the Balkans accounts for the "Roman" character of the major Romance languages. Later European colonial and commercial contacts with parts of the Americas, of Africa, and of Asia readily explain the French, Spanish, and Portuguese spoken in those regions.

The Roman origins of Romance

GENERAL CONSIDERATIONS

Origins and distribution. The name Romance indeed suggests the ultimate connection of these languages with Rome: the English word is derived from an Old French form of Latin *Romanicus*, used in the Middle Ages to designate a vernacular type of Latin speech (as distinct from the more learned form used by clerics) as well as literature written in the vernacular. The fact that the Romance languages share features not found in contemporary Latin textbooks suggests, however, that the version of Latin they continue is not identical with that of Classical Latin as known from literature. Nonetheless, although it is sometimes claimed that the other Italic languages (the Indo-European language group to which Latin belonged, spoken in Italy) did contribute features to Romance, it is fairly certain that it is specifically Latin itself, perhaps in a popular form, that is the precursor of the Romance languages.

By the turn of the 21st century some 920 million people claimed a Romance language as their mother tongue, 300 million people as a second language. To this number may be added the not-inconsiderable number of Romance creole speakers (a creole is a simplified or pidgin form of a language that has become the native language of a community) scattered around the world. French creoles are spoken by more than 9 million people in the West Indies, North America, and islands of the Indian Ocean (*e.g.*, Mauritius, Réunion, Rodrigues Island, the Seychelles); Portuguese creoles in Cape Verde, Guinea-Bissau, São Tomé and Príncipe, India, and Malaysia (probably some 570,000 speakers); and Spanish creoles (including Papiamento and Chavacano) spoken by more than 500,000 people in the West Indies and the Philippines. Many speakers use creole for informal purposes and the standard language for formal occasions. Romance languages are also used formally in some countries where one or more non-Romance languages are used by most speakers for everyday purposes. French, for example, is used alongside Arabic in Tunisia, Morocco, and Algeria; it is an (or the) official language of 17 countries in the continent of Africa and of Madagascar. Portuguese is the official language of Angola, Mozambique, Guinea-Bissau, Cape Verde, and São Tomé and Príncipe; Italian is widely used in Somalia.

Although its influence has waned before the growing

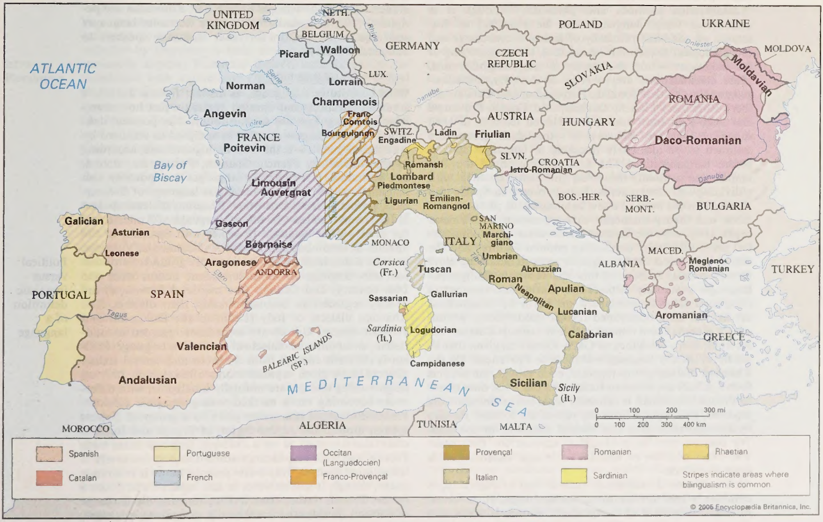


Figure 9: Distribution of Romance languages in Europe.

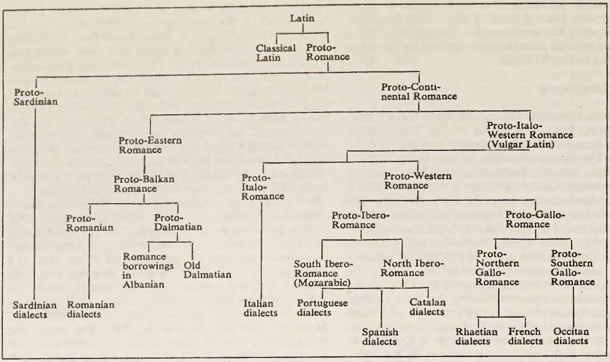


Figure 10: Derivation of Romance languages from Latin.
From R.A. Hall Jr., *Introductory Linguistics*; originally published by Chilton Books, now distributed by Rand McNally & Co.

popularity of English as an international language, French is still widely used today as a second language in many parts of the world. The wealth of French literary tradition, its precisely formulated grammar bequeathed by 17th- and 18th-century grammarians, and the pride of the French in their language may ensure it a lasting importance among languages of the world. By virtue of the vast territories in which Spanish and Portuguese hold sway, those languages will continue to be of prime importance. Even though territorially it has comparatively little extension, the Italian language, associated with Italy's great cultural heritage, is still popular with students.

Classification methods and problems. Though it is quite clear which languages can be classified as Romance, on the basis primarily of lexical (vocabulary) and morphological (structural) similarities, the subgrouping of the languages within the family is less straightforward. Most classifications are, overtly or covertly, historico-geographic—so that Spanish, Portuguese, and Catalan are Ibero-Romance; French, Occitan, and Franco-Provençal are Gallo-Romance; and so on. Shared features in each subgroup that are not seen in other such groups are assumed to be ultimately traceable to languages spoken before Romanization. The first subdivision of the Romance area is usually into West and East Romance, with a dividing line drawn across Italy between La Spezia and Rimini. On the basis of a few heterogeneous phonetic features, one theory maintains that separation into dialects began early, with the Eastern dialect areas (including central and south Italy) developing popular features and the school-influenced Western speech areas maintaining more literary standards. Beyond this, the substrata (indigenous languages eventually displaced by Latin) and superstrata (languages later superimposed on Latin by conquerors) are held to have occasioned further subdivisions. Within such a schema there remain problem cases. Is Catalan, for instance, Ibero-Romance or Gallo-Romance, given that its medieval literary language was close to Provençal? Do the Rhaetian dialects group together, even though the dialects found in Italy are closer to Italian and the Swiss ones closer to French? Sardinian is generally regarded as linguistically separate, its isolation from the rest of the Roman Empire by incorporation into the Vandal kingdom in the mid-5th century providing historical support for the thesis. The exact position of Dalmatian in any classification is open to dispute.

A family-tree classification, such as that of Figure 10, is commonly used for the Romance languages. If, however, historical treatment of one phonetic feature is taken as a classificatory criterion for construction of a tree, results differ. Classified according to the historical development of stressed vowels, French would be grouped with North Italian and Dalmatian but not with Occitan, while Central Italian would be isolated. Classifications that are not based on family trees usually involve ranking languages according to degree of differentiation rather than grouping them; thus, if the Romance languages are compared with Latin, it is seen that by most measures Sardinian and Italian are least differentiated and French most (though in vocabulary Romanian has changed most). By most nonhistorical measures, standard Italian is a "central" language (*i.e.,* it is quite close and often readily intelligible to all other Romance languages), whereas French and Romanian are peripheral (they lack similarity to other Romance languages and require more effort for other Romance speakers to understand them).

LANGUAGES OF THE FAMILY

What constitutes a language, as distinct from a dialect, is a vexing question, and opinion varies on just how many Romance languages are spoken today. The political definition of a language—one that is accepted as standard by a nation or people—is the least ambiguous one; according to this definition, French, Spanish, Portuguese, Italian, and Romanian are certainly languages and possibly also Romansh (since 1996 a semiofficial language of Switzerland, probably related to other Rhaetian dialects spoken in Italy) and Catalan (the official language of Andorra and the joint official language [together with Spanish] of the Spanish autonomous communities of Catalonia, Valencia, and the Balearic Islands). On linguistic grounds Sardinian (not the language of an independent nation since the 14th century) and Occitan (the medieval Provençal) are usually regarded as languages rather than dialects. The Rhaetian dialects of Italy (Ladin in the Dolomites and Friulian around Udine) are sometimes regarded as non-Italian, sometimes as dialects of the Italian language. Sicilian is different enough from northern and central Italian dialects to be given separate status often, but in Italy all neighbouring dialects are mutually intelligible, with differences becoming more marked with geographic distance. Franco-Provençal (the name given to a group of dialects spoken around the Alpine region of France and Italy) is often also assumed to be a different language from both French and Occitan, though some think it is merely a transitional dialect. Only a few persons know it in France today, though it still survives in the Italian Valle d'Aosta (where French, rather than Italian, remains the language of culture). Besides the traditionally recognized languages, there also exist dialects that, given their written tradition

Political versus linguistic definition of a language or the active promotion of their use in writing, demand to have their independence acknowledged: Galician, Asturian, Piedmontese, and Corsican.

Ladino (not to be confused with Ladin), or Judeo-Spanish, is normally regarded not as an independent language but as an archaic form of Castilian Spanish preserving many features of the 15th-century language that was current when the Jews were expelled from Spain. There are possibly about 160,000 speakers, mostly originating in the Balkans and Asia Minor but, after World War II, concentrated in Israel; about 100,000 speakers now reside in Israel, and others live in Turkey and the United States.

Some linguists believe that creoles are often different languages from their metropolitan counterparts; Haitian, for instance, is said to be mutually unintelligible with French. Intelligibility varies so much with the speaker and the hearer, however, that it is difficult to formulate firm criteria on this basis.

Many Romance dialects have virtually ceased to be spoken in the 20th century. Of these, Dalmatian is the most striking, its last known speaker, one Tuone Udaina (Italian Antonio Udina), having been blown up by a land mine in 1898. He was the main source of knowledge for his parents' dialect (that of the island of Veglia [modern Krk], though he was hardly an ideal informant; Vegliot Dalmatian was not his native language, and he had learned it only from listening to his parents' private conversations. Moreover, he had not spoken the language for 20 years at the time he acted as an informant, and he was deaf and toothless as well. Most of the other evidence for Dalmatian derives from documents from Zara (modern Zadar) and Ragusa (modern Dubrovnik) dating to the 13th–16th centuries. It is possible that, apart from isolated pockets, the language was then replaced by Croatian and, to a lesser extent, by Venetian (a dialect of Italian). It is certain, even from scanty evidence, that Dalmatian was a language in its own right, noticeably different from other Romance languages.

On the Istrian Peninsula of Croatia close to the island of Krk, another Romance variety precariously survives with probably fewer than one thousand speakers; known as Istriot, it may be related to Vegliot. Though some scholars connect it with Rhaetian Friulian dialects or with Venetian dialects of Italian, others maintain that it is an independent language. There are no texts except those collected by linguists. A little farther north in the same peninsula, another Romance dialect, Istro-Romanian, is threatened with extinction. Usually classified as a Romanian dialect, it may have been carried to the Istrian Peninsula by Romanians from the northwestern part of the Balkan Peninsula who took refuge from the Turks in the 16th and 17th centuries; it has undergone strong Croatian influence. The first evidence of its existence is a short list of words in a historical work of 1698; there are also collections of folklore texts from the 19th century, but it is otherwise unwritten. Another isolated Romanian dialect that may be nearing extinction is Megleno-Romanian (Meglenitic), from a mountainous region of Macedonia, just west of the Vardar River, on the border between the Republic of Macedonia and Greece. In 1914 there were 13,000 speakers, but many have emigrated to Asia Minor, other parts of what was once Yugoslavia, and Romania, where small pockets survive (they numbered about 5,000 in the late 1990s). The only texts are those transcribed from oral traditions.

Istro-Romanian

Other Romance tongues earlier ceased to be spoken; there is evidence, for instance, of an Ibero-Romance dialect spoken in Arab-occupied Spain until shortly after its reconquest by the Spanish, accomplished at the end of the 15th century. Usually known as Mozarabic, from the Arabic word for 'Arabized person,' or as *'ajamī* ('barbarian language'), it was originally the spoken language of the urban bourgeoisie, who remained Christian while the peasantry generally converted to Islām, but it appears that many Arabs also came to use it as a spoken language. Because most of the evidence, apart from a 15th-century glossary from Granada, is written in Arabic script (which uses no vowel signs), it is difficult to reconstruct the phonology of the language, but it appears to be a very conservative Ibero-Romance dialect. Much of modern information about Mozarabic comes from medical and botanical works that give Mozarabic terms alongside the Arabic. To this was added the discovery of Mozarabic refrains (*kharjahs*) attached to Arabic love ballads (*muwashshaḥs*) of the 11th and 12th centuries; study of these began only in 1948. For much of the Muslim period (beginning in 711), Christians were treated tolerantly and became culturally Arabized. Even after persecution by fanatic Muslim newcomers in the 12th century, the Mozarabs were often in conflict with Westernized "liberators" from the north. Their language died out soon after the Arabs were driven out of Spain at the end of the 15th century, though it is sometimes claimed that Mozarabic has left its mark on the dialects of southern Spain and Portugal.

Other Romance varieties may have developed in peripheral regions of the Roman Empire only to die out under pressure from neighbouring non-Italic languages; these regions are called Romania submersa by specialists. Often these extinct Romance varieties are known from words borrowed into surviving languages; Berber, for instance, bears witness to the long and brilliant Roman period in North Africa that ended in the 7th century AD with Arab invasions, and British Celtic (especially Welsh) retains many traces of what appears to have been a conservative Romance dialect, otherwise eliminated by Anglo-Saxon in the 5th century. Albanian contains so many Romance words that some style it "semi-Romance," and farther north, in what was formerly the Roman province of Pannonia (corresponding to modern western Hungary and parts of eastern Austria, Slovenia, and northern Yugoslavia), Romance speech was probably not dead at the time of the Magyar invasion at the end of the 10th century. Thus, there is reason to believe that Romance dialects may have been spoken at one time over much of southeastern Europe. It is also evident that Romance languages have been retreating south before German for some time, and it is probable that Romance tongues were used in the whole of Switzerland and parts of Bavaria and Austria until roughly the 9th to 10th century.

The once-vast extent of Romance languages

Major languages. *French.* Probably the most internationally significant of the Romance languages, French is used as the official language in 22 countries and as a co-official language in several more (including Belgium, Canada, Haiti, Madagascar, and Switzerland). In France and Corsica more than 51 million individuals use it as their first language; in Canada, more than 7.2 million; in Belgium, more than 3.3 million; in Switzerland (cantons of Neuchâtel, Vaud, Genève, Valais, Fribourg), more than 1.2 million; in Monaco, some 17,000; in Italy, some 100,000; and, in the United States (especially Maine and Louisiana), almost 2 million. Furthermore, more than 5 million Africans (in such countries as Benin, Burkina Faso, Cameroon, Central African Republic, Chad, Congo, Côte d'Ivoire, Djibouti, Guinea, Madagascar, Mali, Niger, Rwanda, Senegal, Togo, and Zaire) and millions of inhabitants of Vietnam, Laos, and Cambodia use it as their principal international language; many creole French speakers, too, use standard French in formal situations.

Standard French is based on the dialect of Paris (in the so-called Île-de-France with its Francien dialect), which assumed importance in the second half of the 12th century; it was basically a north-central dialect with some northern features. Before that, other dialects, especially Norman (which developed in Britain as Anglo-Norman, widely used until the 14th century) and northern dialects (such as Picard), had more prestige, especially in the literary sphere. The legal reform known as the Edict of Villers-Cotterêts (1539), however, established Francien as the only official language (as against both Latin and other dialects) after it had proved to be the most popular written form. From then on, standard French began to replace local dialects, which were officially discouraged until recent times, though the standard language did not spread to popular usage in all regions until well into the 19th century. Dialectal features, which were still admired and cherished by 16th-century writers, were ridiculed in the 17th and 18th centuries, when the grammar and vocabulary of the modern language were standardized and polished to an unprecedented degree.

Linguistic change was more rapid and more drastic in northern France than it was in other European Romance regions, and influence from Latin was comparatively slight (though borrowing of Latin vocabulary has been great since the 14th century). The influence of the Germanic Frankish invaders is often held to account for exotic features in Old French, such as strong stress accent and abundant use of diphthongs and nasal vowels; but by the 15th century the language had begun to change, and a sober (even monotonous) intonation and loss of a stress accent became characteristic. The popularity of French as a first foreign language, in spite of numerous pronunciation difficulties for nearly all foreign speakers, is perhaps as much the result of the precise codification of its grammar, effectuated especially during the 18th century, as of the brilliance of its literature at all periods.

French as a second language

The first document apparently written in French probably dates from 842; known as the Strasbourg Oaths, it is a Romance version of oaths sworn by two of Charlemagne's grandsons. Some claim that the text of this document is thinly disguised Latin constructed after the event to look authentic for political propaganda purposes; others suppose that its Latinizing tendencies reveal the struggle of the scribe with the problems of spelling French as it was spoken at the time. If the language of the Strasbourg Oaths is northern French, it is difficult to ascertain what dialect it represents—some say that of Picard, others Franco-Provençal, and so on.

The second existing text in Old French (with Picard and Walloon features) is a rendering of a short sequence by Prudentius on the life of St. Eulalia, precisely dated (AD 880–882). Two 10th-century texts (the *Passion du Christ* and the *Vie de St. Léger*) seem to mingle northern and southern dialect features, while another (the "Jonas fragment") is obviously from the far north. In the 12th century, the "gem" of the epic poems known as chansons de geste, *La Chanson de Roland*, was written. One of the most beautiful poems of its type in world literature, it evinces certain dialectal characteristics the origins of which are difficult to establish. In the 12th–13th century the Francien dialect became dominant, and it gained the status of literary language because of both the central position of the Île-de-France region and the political and cultural prestige of Paris.

Modern dialects are classified mainly on a geographic basis, and most survive only in rural areas. Walloon, a dialect spoken mainly in Belgium, is something of an exception in that it has had a flourishing dialect literature since approximately 1600. Other dialects are grouped as follows:

Central: Francien, Orléanais, Bourbonnais, Champenois
Northern: Picard, Northern Norman
Eastern: Lorrain, Bourguignon (Burgundian), Franc-Comtois
Western: Norman, Gallo (around the Celtic Breton area), Angevin, Manceau
Southwestern: Poitevin, Saintongeais, Angoumois

Outside France, the French of Canada, originally probably of northwestern dialect type, has developed the most individual features. Although 18th-century Canadian French was regarded as exceptionally "pure" by metropolitan commentators, it began to diverge from Parisian French after 1760 as a consequence of its isolation from the metropolis and of the ever-stronger influence of English. Canadian French is less clearly articulated, with less lip movement and with a more monotonous intonation, than standard French; some change in consonantal sounds occurs (/t/ and /d/ shift to /ts/ and /dz/, respectively, and both *k* and *g* followed by the letters *i* or *e* become palatalized [pronounced with the tongue touching the hard palate, or roof of the mouth]); nasal vowels tend to lose the nasal element; vocabulary and syntax are heavily Anglicized. Although intellectuals turn toward France for cultural inspiration, the pronunciation and usage of standard French is sometimes derided by French Canadians; this may be because their English compatriots are taught Parisian French at school. The French-speaking population of Canada is growing relatively fast, and at present four-fifths of the population of Quebec province use French as their everyday language. The activities of the separatist movement are evidence of the feeling of grievance that many French Canadians still have.

Characteristics of Canadian French

Spanish. Spanish, the Romance language spoken as a first language by the most people in the world, is the (or an) official language of 19 American countries as well as of Spain and Equatorial Guinea. Although many South and Central Americans use native Indian languages as their first language, Spanish is continuing to spread. Estimated numbers of speakers are as follows (in order of numerical importance): Mexico, 85 million; Colombia, 35 million; Argentina, 33 million; Spain, 29 million; Venezuela, 21 million; United States and Peru, 19 million; Chile, 13 million; Cuba, 11 million; Ecuador, 10.8 million; Dominican Republic, 7.3 million; Guatemala, 7 million; El Salvador, 5.9 million; Honduras, 5.5 million; Nicaragua, 4 million; Costa Rica, 3.3 million; Bolivia 3.2 million; Uruguay, 3 million; Puerto Rico and Panama, 2 million; Paraguay, 320,000. (The number of Spanish speakers in Equatorial Guinea is not available.) There are also 160,000 Judeo-Spanish speakers.

The dialect spoken by nearly all Spanish speakers is basically Castilian, and indeed Castellano is still the name used for the language in several American countries. The other languages spoken in Spain include Basque, Catalan (mentioned above), Galician, and Valencian. The now-unchallenged ascendancy of Castilian is the result of the particular circumstances of the Reconquista (the conquest of Moorish Spain by the Christian states of Spain, completed in 1492), with which the language spread to the south. Having established itself in Spain, the Castilian dialect, possibly in its southern, or Andalusian, form, was then exported to the New World during the 16th century.

Standard Castilian is no longer the language of Old Castile, which was regarded as rustic and archaic already in the 15th century, but a modified form developed in Toledo in the 16th and 17th centuries and, more recently, in Madrid. American countries have developed their own standards, differing mainly in phonology (in which they often agree with the southern Spanish dialects) and in vocabulary (in which loanwords from English are more frequent), but differentiation is comparatively slight, and some Americans still regard true Castilian as their model. On the whole, American forms of Spanish are more musical and suave than the Castilian of Madrid, but it is remarkable how little deformation, or creolization, of the language has occurred.

American Spanish versus Castilian Spanish

The first texts in Spanish consist of scattered words glossing two Latin texts of the 10th century, one from Rioja and the other from Castile; the language in the two documents shows few dialect differences. Another document, written about 980, seems to be Leonese in character. The Mozarabic verse forms known as *kharjahs* are the next-oldest surviving texts, but by the middle of the 12th century the famous epic poem *Cantar de mío Cid* ("Song of My Cid") appeared in a language that is basically Castilian. Literary works in Leonese appear until the 14th century and in a conventionalized Aragonese until the 15th century, but Castilian was destined from the first to gain the upper hand, even making an impact on Portuguese, especially during the 15th and early 16th centuries.

Judeo-Spanish is the continuation of an archaic form of Castilian, reflecting the state of the language before 16th-century standardization. The expulsion of the Jews from the Iberian Peninsula in 1492 affected mainly the humbler classes, with the rich preferring "conversion," but the latter often later chose voluntary exile to settle in England and the Netherlands, where their Sephardic tongue precariously survives as a religious language in a few communities. Earlier refugees fled to the Middle East and, once settled, continued to produce learned works in a literary archaic form of Castilian, written in an adapted Hebrew script. This variety, Judeo-Spanish, has two variants: a vernacular (Djudezmo) and Ladino, used in the translations of biblical texts. The former has borrowings from various languages (Balkan languages, French); the latter, more archaic, is more strongly influenced by Hebrew. After further dispersion during and after World War II, Djudezmo is now threatened with extinction, though Ladino survives with a mainly religious function.

Portuguese. Portuguese owes its importance (as the second Romance language [after Spanish] in terms of numbers of speakers) largely to its position as the language of Brazil, where more than 150 million people speak it. In Portugal itself there are about 10 million speakers. The Galician (Gallego, Galego) language of northwestern Spain is historically a Portuguese dialect, though now much influenced by the standard Castilian Spanish; about 2.5 million speakers use Galician as their home language. It is estimated that there are also some 4.6 million Portuguese speakers in Africa (some of whom also use creole) and about 500,000 in the United States.

There are five main Portuguese dialect groups, all mutually intelligible: (1) Northern, or Galician, (2) Central, or Beira, (3) Southern (Estremenho, including Lisbon, Alentejo, and Algarve), (4) Insular, including the dialects of Madeira and the Azores, and (5) Brazilian. Standard Portuguese was developed in the 16th century, basically from the dialects spoken from Lisbon to Coimbra. Brazilian (Brasileiro) differs from the Portuguese spoken in Portugal in several respects, in syntax as well as phonology and vocabulary, but many writers still use an academic metropolitan standard. A creolized form, once widespread in Brazil, seems now to be dying out. A Judeo-Portuguese is attested in 18th-century Amsterdam and Livorno (Leghorn, Italy), but virtually no trace of that dialect remains today.

Galician and the courtly lyric

In the region of northwestern Spain that adjoins Portugal, the Galician dialects lack uniformity and are closer to Spanish. Even in Castile, where standard Spanish (Castilian) originated, Galician was the conventional language of the courtly lyric until roughly 1400, but it lost ground in the 15th century, and Castilian replaced Galician as the official language of Galicia in 1500. Dialect poetry in Galician has flourished from the 18th century, with an upsurge in the 19th century.

Until the 15th century, Portuguese and Galician formed one single linguistic unit, Gallego-Portuguese. The first evidence for the language consists of scattered words in 9th–12th-century Latin texts; continuous documents date from approximately 1192, the date assigned to an extant property agreement between the children of a well-to-do family from the Minho River valley.

Literature began to flourish especially during the 13th and 14th centuries, when the soft Gallego-Portuguese tongue was preferred by courtly lyric poets throughout the Iberian Peninsula except in the Catalan area. In the 16th century, Portugal's golden age, Galician and Portuguese grew further apart, with the consolidation of the standard Portuguese language. From the 16th to the 18th century, Galician was used only as a home language (*i.e.,* as a means of communication within the family). Toward the end of the 18th century, it was revived as a language of culture. Today, with Spanish, it is an official language of the *comunidad autónoma* ("autonomous community") of Galicia.

Italian. Italian is currently spoken by more than 66 million people, of whom the vast majority live in peninsular Italy. It is the official language of Italy, San Marino, and (together with Latin) Vatican City. Italian is also (with German and French) an official language of Switzerland's Ticino *canton,* where it is spoken by some 500,000. Italian is also used as a common language in France (the Alps and Côte d'Azur) and in small communities in Croatia and Slovenia. On the island of Corsica a Tuscan variety of Italian is spoken, though Italian is not the language of culture. Overseas (*e.g.,* in the United States, where it is estimated that there are some 1,500,000 Italian speakers; in Brazil, with about 700,000, and in Argentina, with about 600,000) speakers sometimes do not know the standard language and use only dialect forms. Increasingly, they only rarely know the language of their parents or grandparents.

Speakers of an Italian dialect, even one as superficially different as Sicilian, can with effort understand standard Italian, however, and can even learn it by such means as listening to radio and television programs. For most Italians, first contact with the standard language comes in primary school, in which until recently it was the only dialect used; standard Italian is virtually the only dialect of culture in modern Italy, and with immigration from the south to the industrial north it became the language of intercommunication. Standard Italian is widely used in Somalia, but no longer in Malta. In Libya and Ethiopia, too, its use is dying out.

Standard Italian began to be developed in the 13th and 14th centuries as a literary dialect. At first basically a Florentine dialect stripped of local peculiarities, it has since acquired some characteristics of the dialect of Rome in particular and has always been heavily influenced by Latin. It overlies a wide variety of dialects, which are sometimes considered to represent a fundamental differentiation between northern and southern Italy that dates from Roman times.

The dialect continuum

Today, however, these variant dialects form a continuum of intelligibility, although geographically distant dialects may be radically different. The northern dialects include what are often called the Gallo-Italian dialects (Piedmontese, Lombard, Ligurian, Emilian-Romagnol), in which some linguists discern the influence of a Celtic (Gaulish) substratum (*i.e.,* the traces of a language previously spoken in the region). The other northern group of dialects, spoken in northeastern Italy, is called Venetan (including Venetian, Veronese, Trevisan, and Paduan dialects, etc.). The Istriot dialects, which are spoken on the peninsula now divided between Croatia and Slovenia, with a tiny portion belonging to Italy, are still difficult to classify. They were believed to be variants of Friulian and Dalmatian (which have the same Illyrian substrate), independent varieties of an archaic type of Italian, or independent varieties of other Romance languages. The Tuscan dialects (including those of Corsica) are often held to form a linguistic group of their own, while in the south and east three broad dialect areas are grouped loosely together: (1) the dialects of the Marche (Marchigiano), Umbria, and Rome; (2) Abruzzian, Apulian, Neapolitan, Campanian, and Lucanian; and (3) Calabrian, Salintino, and Sicilian, which are believed by some to be influenced by the Greek once spoken there (which still survives in isolated pockets on the extreme southern portion of the peninsula).

Outside Italy, Italian dialects are heavily influenced by contact with other languages (English in New York; Spanish in Buenos Aires). A pidgin Italian can still be heard in Addis Ababa, Ethiopia, but it has little extension. Relics of a Judeo-Italian (Italkian) survive within Italy (especially in Rome and Livorno); an entire colony of 6,000 Corfu Jews, who used a Venetan dialect as a home language, was exterminated during World War II.

Early texts from Italy are written in dialects of the language that only later became standard Italian. Possibly the very first text is a riddle from Verona, dating from perhaps the 8th century, but its language is Latinized. More surely Italian are some 10th-century documents from Montecassino (testimonies in court—*e.g.,* Placiti [decrees] of Capua, of Sessa, and so on), after which there are three central Italian texts of the 11th century. The first literary work of any length is the Tuscan *Ritmo Laurenziano* ("Laurentian Rhythm") from the end of the 12th century, followed soon by other compositions from the Marches and Montecassino. In the 13th century lyric poetry was first written in a conventionalized Sicilian dialect that influenced later developments in Tuscany.

In modern Italy, although dialects are still the primary spoken form, standard Italian is virtually the only written language.

Romanian. There are about 23,680,000 speakers of Romanian, of whom about 20,500,000 live in Romania, 2,700,000 in Moldova, some 350,000 in Ukraine, and about 40,000 in Yugoslavia and 10,000 in Hungary. There are about 80,000 Romanian speakers in the United States. An additional 500,000 speak Aromanian, or Macedo-Romanian, a group of dialects scattered across Macedonia, Greece, Bulgaria, Albania, Yugoslavia, and Romania.

Standard Romanian

The standard language of Romania is based on a Walachian variety of so-called Daco-Romanian, the majority group of dialects; it was developed in the 17th century mainly by religious writers of the Orthodox church and includes features from a number of dialects, though Bucharest usage provides the current model. Daco-Romanian is fairly homogeneous but shows greater dialectal diversity in the Transylvanian Alps, from which region the language may have spread to the plains. During the Soviet era, the language of Moldova was written in the Cyrillic alphabet, called Moldavian, and held by Soviet scholars to be an independent Romance language. Currently called either Romanian or Moldovan, since 1989 the language has been written in the Roman alphabet. The dialects of Romanian are barely mutually intelligible with the standard, and some can be counted as separate languages; these include Megleno-Romanian (Meglenitic) and Istro-Romanian, both of which are nearly extinct, and the more vigorous Aromanian (Macedo-Romanian). Numbers have probably decreased considerably, but certainly before 1940 Aromanians were often prominent businessmen in their localities. The first known inscription in Aromanian, dated 1731, was found in 1952 at Ardenita, in Albania; texts date to the end of the 18th century, and literary texts have been published in the 19th and 20th centuries (mostly in Bucharest).

The first known Daco-Romanian text is a private letter of Walachian origin dated 1521, though some manuscript translations of religious texts show Transylvanian dialect features and may be earlier. The oldest printed texts are *Evangheliarul slavo-român* (1551–52; "The Slavo-Romanian Gospels") of Sibiu and the works of Deacon Coresi, beginning in 1559. The vast majority of early texts are written in Cyrillic script, the Roman (Latin) alphabet having been officially adopted in 1859 at the time of the union of Walachia and Moldavia. Literature in Romanian began to flourish in the 19th century, when the emerging nation turned toward other Romance countries, especially France, for cultural inspiration. This circumstance had important consequences for the language, triggering the so-called re-Romanization of Romanian.

Minor languages. *Occitan.* Occitan (also called Languedoc or Languedocien) is the modern name given by linguists to the group of dialects spoken by some 1.5 million people in the south of France. All Occitan speakers now use French as their official and cultural language, but their local dialects remain lively and, across most of the area, remarkably homogeneous. The name Occitan derives from the old name Occitanie (formed on the model of Aquitaine) of the area now known as Languedoc. The medieval language is often called *langue d'oc,* which denoted a language using *oc* (from Latin *hoc*) for 'yes' in contrast to *langue d'oïl,* denoting French, and the *si* languages, Spanish and Italian. In the area itself, the names Lemosí (Limousin) and Proensal (Provençal) were formerly used, but today these names are usually considered too localized to designate the whole range of dialects. Members of a vigorous literary movement in the Provence region, however, still prefer to call their language Provençal.

Occitan was rich in poetic literature in the Middle Ages until the north crushed political power in the south (1208–29). The standard language was well established, however, and it did not really succumb before French until the 16th century, while only after the Revolution of 1789 did the French language penetrate into popular use in place of Occitan. In the mid-19th century, a literary Renaissance led by the Félibres, and based on the dialect of the Arles-Avignon region, lent new lustre to Occitan, and a modern standard dialect was established. The most famous figure of this movement was Frédéric Mistral, a Nobel Prize-winning poet. Almost contemporaneously, a similar movement, based in Toulouse, arose and concentrated on problems of linguistic and orthographic standardization to provide a wider base for literary endeavour.

The role of the Félibres

The Occitan dialects have changed comparatively little since the Middle Ages, though now French influence is increasingly evident. Perhaps this influence has helped them to remain more or less mutually intelligible. The main dialect areas are Limousin, in the northwest corner of the Occitan area; Auvergnat, in the north-central region of this area; northeastern Alpine-Provençal; and Languedoc and Provençal, on the west and east of the Mediterranean seacoast, respectively.

Gascon, in the southwest of France, is usually classified as an Occitan dialect, though to most other southerners it is today less readily comprehensible than Catalan. Some scholars claim that it has always been distinct from Occitan, because of the influence of a non-Celtic Aquitanian pre-Roman population. The Roman name of the region, Vasconia (from which the name Gascony derives), suggests the relationship of its original population with the non-Indo-European Basques.

Franco-Provençal. Northeast of the Occitan region, along the French, Swiss, and Italian frontiers, is located a group of dialects that historically have shared most vowel developments with languages to the south and many consonant changes with those to the north. Since 1878, with the work of G.I. Ascoli, claims have been made for the linguistic autonomy of these dialects, usually called Franco-Provençal; at the end of the 20th century it was estimated that somewhat fewer than two million speakers use them (urban speakers are hard to find, and even in the countryside young speakers are few). Dialects are extremely diversified and heavily influenced by French, which has been used extensively in the area since the 13th century. Even during the Middle Ages the language had no standard form, though there are some 12th–13th-century documents in Franco-Provençal. The dialect of Geneva (extinct except in some rural communes) was the official language of the Swiss republic for some time, but none of the other dialects has had official status. Some claim that a section of a manuscript, the so-called Alexander fragment, dating from the 11th–12th century and apparently part of a lost poem, is Franco-Provençal in character, but others maintain that it, like other literary texts from the region, is mainly Provençal with some French features. Since the 16th century there has been local dialect literature, notably in Savoy, Fribourg, and Geneva.

Catalan. Currently spoken by more than 6,500,000 people in Spain and some 210,000 in France, as well as by some 20,000 in Andorra and some 30,000 in Alghero (Sardinia), Catalan has lost little of its former lustre, even though it is no longer as widespread as it was between 1137 and 1749, as the official language of Aragon. Although there is no evidence of dialectalization in the Middle Ages, perhaps because of the standardizing influence of its official use in the Kingdom of Aragon, since the 16th century the dialects of Valencia and the Balearic Isles, especially, have tended to differentiate from the Central (Barcelona) dialect. Nevertheless, some degree of uniformity is preserved in the literary language, which has continued to flourish in spite of the little encouragement received after the Spanish Civil War (1936–39). With the administrative reorganization that started in the late 1970s, Catalonia became a *comunidad autónoma* ("autonomous community"), and Catalan once again gained ascendancy in eastern Spain (see Figure 9).

Historical use of Catalan

The earliest surviving written materials in Catalan—a charter and six sermons—date from the 12th century, with poetry flourishing from the 13th century; before the 13th century Catalan poets wrote in Provençal. The first true Catalan poet was Ramon Llull (*c.* 1232/33–1315/16), and the greatest Catalan poet was Ausias March (1397–1459), a Valencian. The language retained its vigour until the union of the Aragonese and Castilian crowns in 1474 marked the beginning of its decline. After that, mainly grammatical works appeared; the language was to wait for its renaissance until the late 19th century. In 1906 the first Catalan Language Congress attracted 3,000 participants, and in 1907 the Institut d'Estudis Catalans was founded. Yet not until 1944 was there a course in Catalan philology at the University of Barcelona; a chair of Catalan language and literature was founded there in 1961.

It is much disputed whether Catalan is more closely related to Occitan or to the Hispanic languages. Medieval Catalan was so close to Lemosí, the literary dialect of Occitan in southern France, that it is thought by some to have been imported from beyond the Pyrenees in the resettlement of refugees from the Moors.

In more modern times Catalan has, however, grown closer to Aragonese and Castilian, so that its family-tree classification becomes less indicative of the living language. It was occasionally called Llemosí by 19th-century

Catalan revivalists, however, who wished to emphasize its independence from other Iberian tongues by stressing its relation to Occitan. Certainly, by most standards, Catalan merits the distinction of being deemed a language in its own right, and it shows little sign of decline.

Sardinian. Sardinian is currently spoken by more than 1.5 million people, but it has many dialect differences. There is virtually no literature, not even a newspaper, in the language (although satirical journals do appear from time to time). In earlier times the language was probably spoken in Corsica, where a Tuscan dialect of Italian is now used (although French has been Corsica's official language for two centuries). From the 14th to the 17th century, Catalan (at that time the official language of Aragon, which ruled Sardinia) was used extensively, especially for official purposes; a Catalan dialect is still spoken in Alghero. Castilian began to be used in Sardinian official documents in 1600 but did not supplant Catalan in the south of the island until later in the 17th century. Since the early 18th century Sardinia's destiny has been linked with that of the Italian mainland, and Italian is now the official language.

Historical extent of Sardinian

Sardinia was more or less independent from 1016, until the arrival of the Aragonese in 1322, though much influenced by the Genoese and Pisan peoples. The first documents in Sardinian are *condaghi,* legal contracts dating from approximately 1080; in the north of the island, Sardinian was used for such documents until the 17th century. Logudorian (Logudorese) was the central and the most conservative dialect. The northern form of Logudorian provides the basis for a *sardo illustre* (a conventionalized literary language that has been used mainly for folk-based verse). Other dialects of Sardinian include Campidanese (Campidanian), centred around Cagliari in the south, heavily influenced by Catalan and Italian; Sassarese (Sassarian) in the northwest; and Gallurese (Gallurian) in the northeast. It is sometimes said that the latter two dialects are not Sardinian but rather Corsican. Gallurese in particular is related to the dialect of Sartène in Corsica, and it may have been imported into the Gallura region in the 17th and 18th centuries by refugees from Corsican vendettas.

Sardinian is unintelligible to most Italians and gives an acoustic impression more similar to Spanish than Italian. It is clearly and energetically articulated but has always been regarded as barbarous by the soft-speaking Italians; Dante, for instance, said that Sardinians were like monkeys imitating men. It retains its vitality as a "home language," but dialects are so diversified that it is not likely to gain greater prominence.

Rhaetian. The Rhaetian, or Rhaeto-Romanic, dialects derive their conventional name from the ancient Raetics of the Adige area, who, according to classical authors, spoke an Etruscan dialect. In fact, there is nothing to connect Raetic with Rhaetian except geographic location, and some scholars would deny that the different Rhaetian dialects have much in common, though others claim that they are remnants of a once-widespread Germano-Romance tongue. Three isolated regions continue to use Rhaetian.

In Switzerland, Romansh, the standard dialect of Graubünden canton, has been a "national" language, used for cantonal but not federal purposes, since 1938. A referendum in 1996 accorded it semiofficial status. The proportion of Rhaetian speakers in Graubünden fell from two-fifths in 1880 to one-fourth in 1970, with a corresponding increase in the Italian-speaking population. In the mid-1990s speakers of Romansh formed about 0.6 percent of the population of Switzerland. Nevertheless, interest in Romansh remains keen, and there are several Romansh newspapers.

The main Romansh dialects

The main Romansh dialects are usually known as Sursilvan and Sutsilvan, spoken on the western and eastern banks of the Rhine, respectively. Another important Swiss Rhaetian dialect, Engadine, is spoken in the Protestant Inn River valley, east of which there is a German-speaking area that has encroached on former Romance territory since the 16th century. The dialects from the extreme east and west of the Swiss Rhaetian area are mutually intelligible only with difficulty, though each dialect is intelligible to its neighbour.

Sursilvan (spoken around the town of Disentis) has one text dating from the beginning of the 12th century but then nothing else until the work of Gian Travers (1483–1563), a Protestant writer. The Upper Engadine dialect (spoken around Samedan and Saint Moritz) is attested from the 16th century, notably with the Swiss Lutheran Jacob Bifrun's translation of the New Testament. Both dialects have had a flourishing local literature since the 19th century. In many ways the Swiss Rhaetian dialects resemble French, and speakers seem to feel more at home with French than with Italian.

In the Trento-Alto Adige region of northeastern Italy, some 30,000 persons speak Ladin. Some Italian scholars have claimed that it is really an Italian (Veneto-Lombard) dialect. The other main language spoken in this now semiautonomous region, much of which was Austrian until 1919, is German, a non-Romance language. Although sometimes said to be threatened with extinction, Ladin appears to retain its vitality among the mountain peasantry. Newspapers are published in Ladin, and the language is comprehensible without too much difficulty to a student of Romance languages. As it appears that these remote valleys were very sparsely populated until the 1960s, the number of speakers there is likely to have grown. Since the 1940s Ladin has been taught in primary schools in the Gardena and Badia valleys, in different conventionalized dialect forms. Although a Ladin document of the 14th century (from the Venosta Valley to the west of the modern Ladin-speaking region) is known from references, the earliest written material in Ladin is an 18th-century word list of the Badia dialect. There are also a few literary and religious texts.

In Italy, north of Venice, stretching to the Slovenian border on the east and to the Austrian border on the north, its western extent almost reaching the Piave River, is the Friulian dialect area, centred around the city of Udine, with some 700,000 speakers. This dialect is much closer to Italian than are Ladin or Romansh, and it is often claimed to be a Venetian dialect. Venetian proper has gained ground at the expense of Friulian to both the east and west since the 1800s. Friulian retains its vitality in the well-populated, industrialized region, however, and supports a vigorous local literature; its most notable poet was Pieri Zorut (1792–1867). The first written specimen of Friulian (apart from a doubtful 12th-century inscription) is a short text dating to approximately 1300, followed by numerous documents in prose, as well as some poems, up to the end of the 16th century, when a rich poetic tradition began.

Creoles. The French, Spanish, and Portuguese creoles, together with their metropolitan equivalents, share many things in common. Indeed, some scholars regard them as in some sense related, either in sharing an African grammatical base with a superimposed Romance lexicon or in historical derivation from a Portuguese pidgin lingua franca used by colonizers and slavers, with later addition of vocabulary from contact with metropolitan languages such as French and Spanish. Other scholars maintain that the creoles are continuators of French, Spanish, and Portuguese in the same way as these are themselves continuators of Latin but that, under the conditions that attended the slave trade, linguistic change was exceptionally rapid, so that the origins of the creoles are often hardly recognizable.

Origin of the term creole

The word *creole* is first found in Spanish (*criollo;* 1590), meaning a Spaniard born in the colonies. It probably originated in Portuguese, although the word *crioulo* has a later attestation. In reference to language, creole has come to indicate a pidgin or trade language that has become the mother tongue of a population, often black; the circumstance under which this usually happened was that of forcible transplantation for the purpose of enslavement. This further brought about the intermingling of individuals with mutually unintelligible native languages and the imposition of the master's language.

Of Romance creoles used today, French creoles are most widespread. In Haiti, which was settled in the mid-17th century, there are some six million creole speakers, of whom only about 10 percent know French, and French creole is an official language, together with French. The Lesser Antilles (Martinique, Guadeloupe, Dominica, Saint Lucia, Saint Kitts and Nevis, etc.) were colonized in 1635, and many of the islands still use French creoles, even when change of ownership led to the imposition of English as the official language.

French creoles also are used in French Guiana and, though dying out, in the U.S. state of Louisiana; Haitian immigrants also account for a large number of the French creole speakers in the United States. In all, more than seven million speakers use French creoles in the Americas. On islands of the Indian Ocean, too, French creoles are spoken; in Mauritius, which was owned by France from 1715 to 1810, French creole (Morisyen) is spoken as a first language by some 600,000 and retains its hold as a lingua franca, even though English is the official language and though a large part of the population uses an Indo-Aryan or Dravidian language at home. In the Seychelles, which were owned by France from 1768 until 1814, and in Réunion (originally L'Île Bourbon), where French is still the official language, French creoles are still in widespread use (in the Seychelles the language is called Seselwa). Some French-creole speakers claim that creoles from other far-off regions are easily intelligible to them. Others contest this, however, pointing out that the creole used by educated speakers is often heavily larded with standard French on all but very informal occasions. Certainly, the linguist can easily discern similarities, especially in grammatical structure, that make the various French creoles seem more like each other than like standard French.

Portuguese creoles were purportedly once widely used in Asia, though probably more frequently as trade languages than as mother tongues. They survive today in Hong Kong and to some extent in Sri Lanka, Malaysia, and a portion of the west coast of India. In Africa, Portuguese creoles are used by more than 450,000 people in Guinea-Bissau, Senegal, the Cape Verde Islands, and some Gulf of Guinea islands (Annobón in Equatorial Guinea, where it is losing ground to Spanish, and São Tomé and Príncipe, where more than four-fifths of the 125,000 inhabitants speak creole dialects). In Brazil a Portuguese creole was once widespread, extending even to Suriname, where Portuguese Jews and their slaves fled in the 17th century; the creole is now virtually extinct.

Papiamento

Papiamento, spoken by more than 225,000 people on the islands of Aruba, Curaçao, and Bonaire in the Caribbean Sea, is an Iberian creole, with both Portuguese and Spanish influence. Chavacano, a Spanish creole, also survives in the Philippines among descendants of mixed Spanish-Filipino stock.

On the whole, creoles are rarely used for literature, except satirical and comic pieces. Most speakers regard them as "inferior" versions of the standard language and in formal situations try to improve their usage on the model of the standard, though they admit to feeling more relaxed speaking creole.

LATIN AND THE DEVELOPMENT OF THE ROMANCE LANGUAGES

Latin and the protolanguage. Latin is traditionally grouped with Faliscan among the Italic languages, of which the other main member is the Osco-Umbrian group (see also *Italic languages* above). Oscan was the name given by the Romans to a group of dialects spoken by Samnite tribes to the south of Rome. It is well attested in inscriptions and texts for about five centuries before Christ and was used in official documents until approximately 90–89 BC. The absence of great dialectal variations in the texts suggests that they are written in a standardized form, though three alphabets are evident—the local one (derived from Etruscan), the Greek (in the southern cities), and the Latin (in more recent inscriptions). In early times, Umbrian was spoken northeast of Rome, to the east of the Etruscan region, and possibly as far west as the Adriatic Sea at one period. It is attested mainly in one series of texts, the Iguvine Tables (Tabulae Iguvinae), dated from 400 to 90 BC, and it is similar to Oscan. Probably Latin and Osco-Umbrian were not mutually intelligible; some claim they are not closely related genetically but that their common features arose from convergence as a result of contact.

The Roman dialect was originally one of a number of Latinian dialects, of which the most important was Faliscan, the language of Falerii (modern Civita Castellana), the most important Faliscan city, located 32 miles (51 kilometres) north of Rome. The Faliscans were probably a Sabine tribe that early fell under Etruscan domination. The dialect is known mainly from short inscriptions dating to the 3rd and 2nd centuries BC and probably survived until well after the conquest of Falerii by the Romans in 241 BC.

The earliest Latinian text is an inscription on a cloak pin (fibula) of the 6th century BC, from Palestrina (Praeneste). Other Latinian inscriptions show marked differences from Roman Latin, for which there is, however, little evidence before the end of the 3rd century BC. What is certain is that the language changed so rapidly between the 5th century (the date of a mutilated inscription, probably a religious prescription, found in the Roman Forum and of the Twelve Tables, the contents of which are known from later evidence) and the 3rd century BC that older texts were no longer intelligible.

During this period the Romans subjugated their Latin neighbours (by 335 BC), and their language began to establish itself as a standard form, absorbing features from other dialects. The first author of any note was the comic dramatist Plautus (*c.* 254–184 BC), whose language is thought to reflect a spoken idiom, some features of which appear to have survived into Romance.

The language of Plautus

By 265 Rome had conquered Magna Graecia, in the south of the Italian Peninsula, and had begun to absorb Greek literary and cultural ideals. Poetic language was especially influenced by Greek until Latin poetry reached its zenith with Virgil. In the 1st century BC a literary prose developed; it emphasized elegance and clarity and rejected vulgarity and rusticity. Grammatical rules were codified and tightened and vocabulary pruned, and the cult of the harmonious, balanced period held sway in rhetorical circles. With Cicero the prose style of the Golden Age attained its highest point; for the linguist, the distinction Cicero makes between the style of his letters and that of his speeches is especially interesting in that it provides evidence that even educated speech differed from written language. When Cicero uses the *sermo plebeius* ("plebeian speech"), his language is more elliptical, with shorter, less complex sentences and more colourful vocabulary (including plentiful diminutives). It seems obvious that truly popular language differed even more from the elaborate, sophisticated, classical literary idiom; there is evidence that archaic features, banned from literary style, survived in common speech right through to the Romance stage of the language. It is sometimes claimed that the language of the Roman historian and politician Sallust (*c.* 86–35/34 BC) approximated popular usage, but it is more probable that his archaizing style derives more from conscious imitation of old Roman poetry. The Roman "judge of elegance" Petronius Arbiter (d. AD 66) is thought to be imitating vulgar speech in the language of the character Trimalchio in the *Cena Trimalchionis* ("Banquet of Trimalchio") section (chapters 26–78) of his *Satyricon.*

Notable characteristics of Classical Latin. *The sound of Classical Latin.* Evidence for pronunciation of Classical Latin is often difficult to interpret. Orthography is conventionalized, and grammarians' comments lack clarity, so that to a considerable extent it is necessary to extrapolate from later developments in Romance in order to describe it.

The most important of the ambiguities bears on Latin intonation and accentuation. The way in which vowels developed in prehistoric Latin suggests the possibility of a stress accent on the first syllable of each word; in later times, however, the accent fell on the penultimate syllable or, when this had "light" quantity, on the antepenultimate. The nature of this accent is hotly disputed: contemporary grammarians seem to suggest it was a musical, tonal accent and not a stress accent. Some scholars claim, however, that Latin grammarians were merely slavishly imitating their Greek counterparts and that the linking of the Latin accent with syllable vowel length makes it unlikely that such an accent was tonal. Probably it was a light stress accent that was normally accompanied by a rise in pitch; in later Latin, evidence suggests that the stress became heavier.

The system of syllable quantity

The system of syllable quantity, connected with that of vowel length, must have given Classical Latin distinctive acoustic character. Broadly speaking, a "light" syllable ended in a short vowel and a "heavy" syllable in a long vowel (or diphthong) or a consonant. The distinction must have been reflected to some extent in Late Latin or early Romance, for, even after the system of vowel length was lost, light, or "open," syllables often developed in a different way from heavy, or "closed," syllables.

Because the system of vowel length was lost after the Classical period, it is not known with any certainty how vowels were pronounced at that period; but, because of later developments in Romance, the assumption is that the vowel-length distinctions were also associated with qualitative differences, in that short vowels were more open, or lax, than long vowels. Standard orthography did not distinguish between long and short vowels, although in early times various devices were tried to remedy this. At the end of the Roman Republic a so-called apex (one form looked somewhat like a hamza [ʾ]) often was used to mark the long vowel, but this mark was replaced in imperial times by an acute accent (ʹ). In Classical Latin the length system was an essential feature of verse, even popular verse, and mistakes in vowel length were regarded as barbarous. In later times, however, many poets were obviously unable to conform to the demands of classical prosody and were criticized for allowing accent to override length distinctions.

Besides the long vowels *ā, ē, ī, ō, ū* and the short vowels *ă, ĕ, ĭ, ŏ, ŭ* educated speech during the Classical period also used a front rounded vowel, a sound taken from Greek upsilon and pronounced rather like French *u* (symbolized by y in the International Phonetic Alphabet—IPA) in words borrowed from Greek; in popular speech this was probably pronounced like Latin *ŭ*, though in later times *ī* was sometimes substituted. A neutral vowel was probably used in some unaccented syllables and was written *u* or *i* (*optumus, optimus* 'best'), but the latter rendering became standard. A long *ē*, from earlier *ei*, had probably completely merged with *ī* by the Classical period. Classical pronunciation also used some diphthongs pronounced by educated Romans much as they are spelled, especially *ae* (earlier *ai*), pronounced perhaps as an open *ē* in rustic speech, *au* (rustic open *ō*), and *oe* (earlier *oi*, Late Latin *ē*).

The consonant system. The Classical Latin consonant system probably included a series of labial sounds (produced with the lips) /p b m f/ and probably /w/; a dental or alveolar series (produced with the tongue against the front teeth or the alveolar ridge behind the upper front teeth) /t d n s l/ and possibly /r/; a velar series (produced with the tongue approaching or contacting the velum or soft palate) /k g/ and perhaps /ŋ/; and a labiovelar series (pronounced with the lips rounded) /kʷ gʷ/. The /k/ sound was written *c*, and the /kʷ/ and /gʷ/ were written *qu* and *gu*, respectively.

Of these, /kʷ/ and /gʷ/ were probably single labialized velar consonants, not clusters, as they do not make for a heavy syllable; /gʷ/ occurs only after /n/, so only guesses can be made about its single consonant status. The sound represented by *ng* (pronounced as in English *sing* and represented in the IPA by /ŋ/), written *ng* or *gn*, may not have had phonemic status (in spite of the pair *annus/agnus* 'year'/'lamb,' in which /ŋ/ may be regarded as a positional variant of /g/). The Latin letter *f* probably represented by Classical times a labiodental sound pronounced with the lower lip touching the upper front teeth like its English equivalent, but earlier it may have been a bilabial (pronounced with the two lips touching or approaching one another). The so-called consonantal *i* and *u* were probably not true consonants but frictionless semivowels; Romance evidence suggests that they later became a palatal fricative, /j/ (pronounced with the tongue touching or approaching

The Latin letter *f*

the hard palate and with incomplete closure) and a bilabial fricative, /β/ (pronounced with vibration of the lips and incomplete closure), but there is no suggestion of this during the Classical period. Some Romance scholars suggest that Latin *s* had a pronunciation like that of *z* in modern Castilian (with the tip, rather than the blade, raised behind the teeth, giving a lisping impression); in early Latin it was often weakened in final position, a feature that also characterizes eastern Romance languages. The *r* was probably a tongue trill during the Classical period, but there is earlier evidence that in some positions it may have been a fricative or a flap. There were two sorts of *l*, velar and palatal ("soft," when followed by *i*).

The nasal consonants were probably weakly articulated in some positions, especially medially before *s* and in final position; probably their medial or final position resulted in mere nasalization of the preceding vowel.

In addition to the consonants shown, educated Roman speakers probably used a series of voiceless aspirated stops, written *ph, th, ch,* originally borrowed from Greek words but also occurring in native words (*pulcher* 'beautiful,' *lachrima* 'tears,' *triumphus* 'triumph,' etc.) from the end of the 2nd century BC.

Another nonvocalic sound, /h/, was pronounced only by educated speakers even in the Classical period, and references to its loss in vulgar speech are frequent.

Consonants written double in the Classical period were probably so pronounced (a distinction was made, for instance, between *anus* 'old woman' and *annus* 'year'). When consonantal *i* appeared intervocalically, it was always doubled in speech. Before the 2nd century BC, consonant gemination (doubling of sounds) was not shown in orthography but was probably current in speech. The eastern Romance languages, on the whole, retained Latin double consonants (as in Italian), whereas the western languages often simplified them.

Morphology and syntax. Latin reduced the number of Indo-European noun cases from eight to six by incorporating the sociative-instrumental (indicating means or agency) and, apart from isolated forms, the locative (indicating place or place where) into the ablative case (originally indicating the relations of separation and source). The dual number was lost, and a fifth noun declension was developed from a heterogeneous collection of nouns. Probably before the Romance period the number of cases was further reduced (there were two in Old French—nominative, used for the subject of a verb, and oblique, used for all other functions—and Romanian today has two, nominative-accusative, used for the subject and the direct object of a verb, and genitive-dative, used to indicate possession and the indirect object of a verb), and words of the fourth and fifth declension were absorbed into the other three or lost.

Among verb forms, the Indo-European aorist (indicating simple occurrence of an action without reference to duration or completion) and perfect (indicating an action or state completed at the time of utterance or at a time spoken of) combined, and the conjunctive (expressing ideas contrary to fact) and optative (expressing a wish or hope) merged to form the subjunctive mood. New tense forms that developed were the future in *-bō* and the imperfect in *-bam;* a passive in *-r,* also found in Celtic and Tocharian, was also developed. New compound passive tenses were formed with the perfect participle and *esse* 'to be' (*e.g., est oneratus* 'he, she, it was burdened')—such compound tenses developed further in Romance. In general, the morphology of the Classical period was codified and fluctuating forms rigidly fixed. In syntax, too, earlier freedom was restricted; thus, the use of the accusative and infinitive in *oratio obliqua* ("indirect discourse") became obligatory, and fine discrimination was required in the use of the subjunctive. Where earlier writers might have used prepositional phrases, Classical authors preferred bare nominal-case forms as terser and more exact. Complex sentences with subtle use of distinctive conjunctions were a feature of the Classical language, and effective play was made with the possibilities offered by flexible word order.

New tense forms

The postclassical period. In the postclassical era, Ciceronian style came to be regarded as laboured and boring, and an epigrammatic, compressed style was preferred by such writers as Seneca and Tacitus. Contemporaneously and a little later, florid, exuberant writing—often called African—came into fashion, exemplified especially by Apuleius (2nd century AD). Imitation of classical and post-classical models continued even into the 6th century, and there seems to have been continuity of literary tradition for some time after the fall of the Western Roman Empire.

The growth of the empire spread Roman culture across much of Europe and North Africa. In all areas, even the outposts, it was not only the rough language of the legions that penetrated but also, it seems, the fine subtleties of Virgilian verse and Ciceronian prose. Research in the late 20th century suggests that in Britain, for instance, Romanization was more widespread and more profound than hitherto suspected and that well-to-do Britons in the colonized region were thoroughly imbued with Roman values. How far these trickled down to the common people is difficult to tell; because Latin died out in Britain, it is often thought that it had been used only by the elite, but some suggest that it was a result of wholesale slaughter of the Roman British. It is, however, more likely that the pattern of Anglo-Saxon settlements was not in conflict with the Romano-Celtic and that the latter were gradually absorbed into the new society.

Vulgar Latin. Later Latin (from the 3rd century AD onward) is often called Vulgar Latin—a confusing term in that it can designate the popular Latin of all periods and is sometimes also used for so-called Proto-Romance (*roman commun*), a theoretical construct based on consistent similarities among all or most Romance languages. All three senses of the term Vulgar Latin in fact share common features but, given their different theoretical status, can hardly be called identical or even comparable. When Christianity was officially adopted by the Roman Empire (4th century), Vulgar Latin elements were diffused through certain religious texts. Its "vulgarisms" often called forth apologies from Christian authors, whose false humility seems akin to pride in that they did not succumb to the frivolities of pagan literary style.

Texts in Vulgar Latin

Aside from the numerous inscriptions found throughout the empire, there is no shortage of texts in Vulgar Latin. One of the first is the so-called *Appendix Probi* (3rd–4th centuries AD; "Appendix to Probus[' Grammar]"), which lists correct and incorrect forms of 227 words, probably as an orthographic aid to scribes. This work illustrates some phonological changes that may have already occurred in the spoken language (*e.g.,* loss of unstressed penultimate syllables and loss of final *m*). The Vulgate, St. Jerome's translation of the Bible (AD 385–404), and some of the works of St. Augustine (AD 354–430) are among Christian works written in Vulgar Latin. Particularly amusing and also linguistically instructive is the so-called *Peregrinatio Etheriae* ("Pilgrimage of Etheria"), also called *Itinerarum Egeriae* ("Travel of Egeria"), written probably in the 4th century by a nun, describing her visit to the Holy Land. Medical and grammatical works also abound from approximately AD 400 to the 7th century (among the writers were the provincials Cosentius, from Gaul; Virgilius Maro, from southern Gaul; and St. Isidore of Seville, from Spain).

Some of the characteristics of Vulgar Latin recall popular features of classical and preclassical times and foreshadow Romance developments. In vocabulary, especially, many of the sober classical words are rejected in favour of more colourful popular terms, especially derivatives and diminutives: thus, *portare* 'to carry' (French *porter,* Italian *portare,* etc.) is preferred to *ferre*; *cantare* 'to sing again and again' (French *chanter,* Spanish and Portuguese *cantar,* etc.) to *canere*; *vetulus* 'little old man' (Romanian *vechi,* Italian *vecchio,* French *vieux,* etc.) to *vetus.* In grammar, synthetic constructions typical of Classical Latin are often replaced by analytic; thus, the use of prepositions often makes case endings superfluous. *Ad regem* for *regi* 'to the king,' for instance, or anomalous morphological forms are simplified and rationalized (*e.g., plus,* or *magis, sanus* for *sanior* 'healthier'). Shorter, simpler sentences are preferred, and word order tends to become less flexible.

The most copious evidence for Vulgar Latin is in the realm of phonology, though interpretation of the evidence is often open to dispute, consisting as it does of the confused descriptions of grammarians and the misspellings of bewildered scribes. Much of the evidence points to a strengthening of stress accent during the Late period, leading to the shortening and swallowing of unaccented syllables: thus, *viridem* 'green' becomes *virdem* (*verde* in several Romance languages); *vinea* 'vine' becomes *vinia* (French *vigne,* Spanish *viña* 'vineyard,' etc.).

Among other phonological features of Vulgar Latin, probably the most striking is the loss of the system of long and short vowels. On the whole, long vowels became tense and short vowels lax, resulting in a wholesale change in the rhythm of the language. In the texts there is evidence of the confusion of *ĭ* and *ē* and of *ŭ* and *ō* that has occurred in the western Romance languages. It is to be remembered that even popular Latin verse used measures of vowel length, and there is no evidence to suggest that vowel-length distinctions were lost in vulgar preclassical speech.

Loss of word-final *m*

An archaic feature that does recur in Vulgar Latin is the loss of word-final *m,* of which virtually no trace remains in Romance. It is possible, however, that the written letter of Classical Latin was no more than an orthographic convention for a nasal twang: in scanning Latin verse, the *-m* is always run in (elided) before an initial vowel. Reduction of the diphthongs /ae/ (to /ɛ/) and /au/ (to /ɔ/) seems also to be a popular and dialectal feature reflected in Vulgar Latin texts; in the latter case, however, the Romance languages do not support the hypothesis that the diphthong was reduced early, for it remains in Old Provençal and in Romanian and, probably, in early Old French.

The emergence of Romance. In the European lands in which Romance languages are still spoken, it is of course certain that, at some point, Latin in some form was the normal language of most strata. Whether, however, the Romance languages continue rough peasant dialects of Latin or the usage of more cultured urban communities is open to question. There are those who maintain that the Latin used in each area differentiated as soon as local populations adopted the conqueror's language for any purpose. According to this belief, dialects of Latin result from divergent developments, either through innovations in restricted areas or through the geographically restricted preservation of some features. It is obvious that Latin usage must have differed over a wide area, but the differences may have been merely phonetic and lexical variations—regional accents and usage—not affecting mutual intelligibility; on the other hand, they may have been profound enough to form the basis of further differentiation when administrative unity was lost. The latter hypothesis would suggest a long period of bilingualism (up to perhaps 500 years), as linguistic interference between languages in contact rarely outlives the bilingual stage. Virtually nothing is known about the status of the indigenous languages during the imperial period, and only vague contemporary references can be found to linguistic differences within the empire. It seems odd that not one of the numerous Latin grammarians should have referred to well-known linguistic facts, but the absence of evidence does not justify the assertion that there was no real diversification during the imperial era. Historical parallels are lacking—although the British Empire, for example, did export English to widely different lands, it lasted a comparatively short time, and its linguistic contribution was backed by modern communications media, besides being to some extent negated by nationalist feeling.

What is certain is that, even if popular usage within the Roman Empire showed great diversification, it was overlaid by a standard written language that preserved a good degree of uniformity until well after the administrative collapse of the empire. As far as the speakers were concerned, they apparently thought they were using Latin, though they were often conscious that their language was, through sheer ignorance, not quite as it should be. Not until about the 8th or 9th century—later in some parts—did it strike them that Classical Latin was perceptibly a different language, rather than merely a more polished, cultured version of their own.

The language of religion and culture. With the spread of Christianity, Latin penetrated to new lands, and it was perhaps the cultivation of Latin in a "pure" form in Ireland, whence it was exported to England, that paved the way for an 8th-century reform of the language by Charlemagne. Conscious that current Latin usage was falling short of Classical Latin standards, Charlemagne invited Alcuin of York, a scholar and grammarian, to his court at Aix-la-Chapelle (Aachen); there Alcuin remained from 782 to 796, inspiring and guiding an intellectual renaissance. It was perhaps as a result of the revival of so-called purer Latin that vernacular texts began to appear, for it now became obvious that the vernacular and Latin were not the same language. Thus, in 813, just before Charlemagne's death, the Council of Tours decreed that sermons should be delivered in *rusticam Romanam linguam* ("in the rustic Roman language") to make them intelligible to the congregation.

Charlemagne's language reform

Latin has remained the official language of the Roman Catholic church and as such has been in constant use by most Romance speakers; it was only within the last half of the 20th century that church services began to be conducted in the vernacular. As the language of science and scholarship, Latin held sway until the 16th century, when, under the influence of the Reformation, nascent nationalism, and the invention of the printing press, it began to be replaced by modern languages. Nevertheless, in the West, along with the knowledge of Greek, the knowledge of Latin has remained a mark of the educated person throughout the centuries, although in the mid-20th century the teaching of classical languages in schools declined significantly.

Latin in non-Romance languages. The prestige of Rome was such that Latin borrowings are to be found in virtually all European languages, as well as in the Berber languages of North Africa, which preserve a number of words, mainly agricultural terms, lost elsewhere. Basque has borrowed a good number of words, mainly from administrative, commercial, and military spheres, but it is difficult in some cases to determine whether the terms were later borrowings from Spanish, rather than from Latin. This uncertainty is not present in the case of the 800 Latin words found in three Celtic languages (Welsh, Cornish, and Breton)—words drawn from a wide sphere of activities. In the Germanic languages, borrowed Latin words principally involve trade and often reflect archaic forms. The very large number of Latin words in Albanian form part of the basic vocabulary of the language (including kinship terms) and cover such spheres as religion, although some of them may have been later borrowings from Romanian. In some cases the Latin words found in Albanian have survived in no other part of the former Roman Empire. Greek and Slavic languages have comparatively few Latin words, many of them administrative or commercial in character.

The development of the Romance languages. The question of when Latin ended and Romance began, which has occupied scholars in the past, is largely a problem created by terminology. In some senses, modern Romance languages are regional varieties of one uniform set of speech patterns that resembles the Vulgar Latin of attested texts fairly closely—indeed, the analyses of generative phonologists make the modern "underlying forms" (as distinct from their phonetic representation in speech) look almost identical with the reconstructed ancestor of the Romance languages, Proto-Romance. On the other hand, contemporary speakers are conscious that they are speaking a "different language" from their neighbours, even though they may understand a good deal of their neighbours' discourse. Perhaps the speaker's consciousness is the best measure of divergence; when, one may ask, did Romance speakers realize that they were not using Latin in their everyday speech? Some scholars suggest that the realization must have struck sometime in the 5th century, when barbarians were streaming into the Roman Empire and, supposedly, hindering communication. Others prefer to rely on positive textual evidence, indicative of efforts to make up a written form of Romance distinct from Latin. Such evidence begins to appear only in the 9th century, first in northern France and then in Spain and Italy. The reforms of Charlemagne, reestablishing more classical standards in written Latin, may have been at once cause and result of the development of conventional written forms for vernacular Romance. Perhaps it was also the emergence of a new type of social organization, feudalism, that had linguistic effects as a result of the splitting of the open society of Roman tradition into small closed territorial units.

Early divergence

Romance glosses to Latin texts. From the 7th century onward, consciousness of linguistic change was strong enough to prompt scribes to gloss little-known words in earlier Latin texts with more familiar terms. Though the glosses often reflect Romance forms, however, they are usually given in a Latinate form, and one gains the impression of a few superficial adjustments to archaic but fundamentally comprehensible texts. The best-known set of glosses—to the Vulgate Bible of St. Jerome—formerly belonged to the abbey of Reichenau, on an island in Lake Constance, Germany, and probably dates from the 8th century. The vocabulary of the Reichenau glosses appears to be French in flavour (*e.g., arenam* 'sand' glossed by *sabulo,* French *sable; vespertiliones* 'bats,' by *calvas sorices,* French *chauvesouris*), and some words of Frankish origin appear (*e.g., scabrones* 'beetles' is glossed by *wapces* 'wasps,' *respectant* 'they look about' by *rewardant*). Some of the glosses evince more widely spread innovations: *emit* 'has bought' is glossed by *comparauit* (Romanian *cumpăra,* Spanish *comprar*), *fissura* 'crack' by *crepatura* (Romanian *crăpătură,* Old French *creveure*). The glosses provide some evidence of morphological simplification (*e.g., saniore* 'healthier' is glossed by *plus sano* 'more healthy' and *cecinit* 'he sang' by *cantavit*), but for the most part only lexical items merit comment. Another well-known glossary, known as the Kassel (or Cassel) glosses, probably dates from the very early 9th century. It gives Latin equivalents of German (Bavarian) words and phrases and provides evidence of lexical and phonetic differentiation within Latin that permits scholars to localize the work as probably French or Rhaetian (*e.g., mantun* 'chin,' as compared with modern French *menton*). Although orthographically eccentric, however, the text is obviously meant to represent Latin, not a Romance tongue; when phrases rather than isolated words are glossed, the Latin is often very close to classical models.

The beginning of Romance literature. Later in the 9th century (with the Strasbourg Oaths, possibly, and more clearly in the Eulalia poem), deliberate attempts were made to write vernacular Romance, though the resources of the Latin alphabet were not wholly adequate to the task. That northern French texts were the first to appear is not surprising, for in that region Latin had changed more radically than elsewhere. By the 10th century the need to couch legal documents in more readily comprehensible vernacular, rather than Latin, was felt in other regions. Vernacular literature did not really get under way, however, until approximately the 12th century, when the arts flourished throughout western Europe. Rhaeto-Romance and Romanian, however, had to wait for the Reformation period to take on literary form.

Northern French

There was a good deal of cross-fertilization between Romance literary languages during the period of development of medieval poetry; the example of the Provençal lyric especially left its mark on all vernacular literatures, and borrowing of lexical items from one language to another was abundant. The 13th century saw some shift of linguistic influence from southern to northern France and from Sicily to Tuscany, toward the politically and economically more powerful regions. Portuguese and Catalan developed flourishing literatures somewhat later, taking over some of the traditions of the badly battered southern French region and dominating the literary scene of the Iberian Peninsula. French was fast losing its hold in England, which, a century earlier, had boasted a rich Anglo-Norman literature, and within France the central Parisian dialect began to dominate. In Italy, the Florentine dialect was showing signs of rising to prominence and providing the base for a literary standard.

The rediscovery of classical literature and art, first in Italy and then in other Romance regions, had some considerable effect on the languages in the shape of extensive borrowing from Latin and Greek and, often, conscious attempts to model grammatical constructions in the vernacular on Classical Latin. The Italian standard language, in particular, owes much to the influence of Latin, which it resembles more closely than do the spoken dialects. French, except in the 16th century, was influenced grammatically less by Latin, but from the 14th century onward the habit of preferring words with a quasi-Latin shape to inherited forms became well established, so that much of the French vocabulary has a "learned" appearance. The trickle of Latinisms into Spanish became a flood in the 15th century, and, though Spanish has been more reluctant than French to reject old words, these Latinisms form a considerable proportion of the modern lexicon.

The standardization of the Romance languages. It was in Italy that the "question of the language" first became a matter of hot dispute. Dante himself made an important contribution to the debate on what should constitute a *volgare illustre* (an "illustrious popular speech") capable of rivaling Latin for literary and scholarly purposes. Controversy did not reach its peak, however, until the 16th century. In the Spain of 1492 the completion of the reconquest of Spain from the Arabs and the so-called discovery of America were accompanied by the appearance of Antonio de Nebrija's *Gramática Castellana* ("Grammar of the Castilian Language"), which argues the need for an ennobled language fit for imperial exportation. In 16th-century France, with the Renaissance backed by the Reformation and the advent of printing, French really assumed the remaining scholarly, scientific, and religious functions of Latin, and efforts were made to put together a worthy national language from dialect and Latin sources. The choice of standard was not made definitively, however, until the late 17th century, when, with political power and social influence centred exclusively at the royal court, the only acceptable usage became that of the court. It would seem that social acceptance and advancement were inextricably bound up with correct behaviour, especially linguistic behaviour, so that the well-to-do bourgeoisie set out to ape the speech habits of their "betters"—hence the popularity of works describing *le bon usage* 'good usage.' The influence of French, resplendent with the achievements of the French dramatic poet Racine and of Louis XIV, was destined to remain dominant within the Romance languages; the Golden Age of Spain and Portugal had already passed, and Italy was going through a period of comparative stagnation.

The dimin-ishing role of Latin

The French grammarians of the 18th century had lasting effect on all the Romance standards, concerned as they were with maintaining "purity," eliminating "vulgarity," and strictly codifying usage, often more in accord with logical than linguistic considerations. The belief that correct language is not a birthright but a tool to be carefully fashioned and skillfully handled, that conscious effort was required to allow it to mirror thought with the minimum of distortion, is one that has persisted in Romance and that still has important effects on educational practice. To many English speakers it seems ludicrous that the criterion of competence in a language should be strict adherence to grammar-book rules rather than nativelike performance, but in Romance countries a foreigner is often frowned upon if he permits himself the "negligence" of native usage, rather than the more stilted correct expression.

LINGUISTIC CHARACTERISTICS OF THE ROMANCE LANGUAGES

As a group, the Romance languages share many characteristics. In comparison with Germanic languages, for instance, they seem musical and mellifluous—probably because of the relatively greater importance of vowels than consonants. On the whole, the vowels are clear and bell-like and articulation energetic and precise, though Portuguese and Romanian convey a more muted acoustic impression. Foreigners often think that Romance speech is particularly rapid and voluble, no doubt because individual words receive only light stress (or, in French, no stress), and elision, the running of words into each other within stress groups, is common. Romanian is something of an exception in that speech tempo is comparatively slow. Intonation patterns, surface manifestations of nonlexical meaning, such as interrogation, exclamation, scorn, surprise, and so forth, seem to some to denote excitability and emotional expressiveness in the speakers. Northern French is comparatively sober, with typically about a one-octave range in intonation, but Italian seems to be sung, with sinuous pitch movement over two octaves, and Castilian jumps jerkily and up and down over about an octave and a third.

Grammatically, the modern languages have retained to a greater or lesser extent some of the synthetic character of Latin, principally in the verb, but in Romanian also in the noun. French, since about the 14th century, has undergone the most radical changes in grammatical typology, so that much greater reliance is placed on word order and intonation to convey sentence meaning than on morphological form. Other languages allow a little more flexibility of word order but far less than does Classical Latin.

Retention of the synthetic character of Latin

Dominant purist grammarians have always opposed influence from foreign languages and reproved their fellows for sullying their language with lavish borrowing (at present primarily from English), but they have never been able to stem the flood of neologisms. French vocabulary, particularly, has always been receptive to change and has been as quick to lose old words as to adopt new. Codification of grammar, on the other hand, has had a permanent effect on the stability of the standard languages, even feeding back into spoken usage via the education system. Acceptance of the most minor changes follows long debate and deliberation and requires governmental edicts that decree what can be marked as correct in all-important examinations. Curiously enough, this rigidity and consequent self-confidence have resulted in greater teachability, so that standards of correctness of, for instance, French among Africans or Spanish among American Indians are remarkably high. The moves toward codification were, indeed, originally linked to a desire to give the languages international importance, and language teaching, in the Romance ethos, is indissolubly linked to the diffusion of cultural and moral values.

Typology. As stated previously, the most "central" Romance language is standard Italian, which has retained and even readopted many Latin characteristics. In some ways its morphology lacks the elegance and efficiency of Castilian, which has most ruthlessly eliminated anomalies during the modern period; there are signs in Italian of historical inertia, a harking back to a glorious past, that has hindered popular development. Romanian remains closest in grammatical type to Latin, though its noun-declension system, based on the placement of the definite article after the noun, and its frequent use of the subjunctive mood may owe much to its Balkan neighbours (or to an earlier linguistic substratum). Its vocabulary has incorporated so many Slavic and Turkish words, however, that it often appears less typical of the Romance languages than the rest. French, by any standard, has diverged most—radical phonetic changes that transformed the outward appearance of the language must have preceded the earliest surviving (9th-century) texts. Such changes are usually ascribed to Celtic and Frankish influence. Another wave of change, with loss of word accent and of many morphological markers, probably dates from the 15th century, but it is difficult to find external motivation for these phenomena. Occitan and Catalan are conservative in character; the long persistence of Roman schools in South Gaul is often seen as the cause of stability there. Spanish and Portuguese are similar enough to lead some scholars to assign their shared characteristics to the influence of an Iberian substratum and a Moorish superstratum. Castilian's forceful character and receptivity to grammatical innovation contrast sharply with Portuguese softness and its inertia in retaining morphological oddities, however. Rhaeto-Romance and Dalmatian peculiarities can most easily be connected with the impact of other languages (mainly German, Italian, and Serbo-Croatian), while Sardinian is often regarded as an extremely conservative, peasant language, some dialects of which have been penetrated by features from Italian and Spanish.

Phonology. Some important phonological develop-

The dating of developments

ments, such as the loss of the system of contrasting vowel lengths and the strengthening of the stress accent, must have occurred during the Vulgar Latin period, while some degree of unity still existed among the various Romance dialects. Certain other changes shared by the Western Romance languages, especially the collapse of *ē* /i/ and *ĭ* /ɪ/, might have postdated the linguistic separation of Sardinia and parts of southern Italy from the other areas, while the distinct development of *ō* /p/ and *ŭ* /u/ in Romanian and Vegliot suggests a split between Eastern and Western Romance at a later date.

Vowels. Everywhere, unaccented vowels have had a different history from accented, and in some languages they have so weakened as to disappear altogether in certain positions. At the end of a word, for instance, even *-a*, the most sonorous of the vowels, has weakened to a neutral vowel in Romanian, Portuguese, and some Catalan and Rhaetian dialects—in some French dialects it is still pronounced as a neutral vowel sound (such as the second vowel in English *alphabet*), but it has been lost completely in the standard language. Final *-o*, from Latin *-ō* or *ŭ*, was lost very early in French, Occitan, Catalan, and Rhaetian and remains only before an article following the word in Romanian; in Portuguese and Romanian it is closed to a /u/ (pronounced as the *u* in English *lunar*). Final *-e* is even more evanescent, regularly remaining as a full vowel only in parts of central and southern Italy and Sardinia.

Under the main stress accent of the word, Latin vowels have often become diphthongs in Romance, perhaps as a result of lengthening under heavy stress or as a consequence of the raising influence of following high vowels (a process known as breaking, similar in action to German umlaut). The vowels most affected are the "open" *e* sound /ɛ/ (as in *met*), from Latin *ĕ*, and to a lesser extent the "open" *o* sound /ɔ/ (pronounced like the vowel sound in *law* in many American English dialects and like the *o* in British English *ingot*), from Latin *ŏ*, while high close vowels *i* /i/ and *u* /u/ are virtually untouched. Transformation of short *e* /ɛ/ to a diphthong (usually /ye/ as in English *yet*) is so common that some believe it occurred during the Vulgar Latin period. The conditions of this process (and similar ones) vary, however; in some languages (notably French and Italian) it happens only in open syllables (*i.e.*, those ending in a vowel in Vulgar Latin), whereas Romanian, Vegliot, Spanish, and perhaps Rhaetian show similar developments in all accented syllables. Portuguese possibly did not join in the diphthong-forming process at all, though, as in Occitan, Catalan, Sardinian, and some Italian dialects, the short *e-* and *o-* sounds may at one time have developed into diphthongs under the influence of a following high vowel (*i* or *u*), later to be reduced once more to a single vowel. Table 10 illustrates treatment of stressed Latin *ĕ* and *ŏ* in different languages.

Latin *ō* (and *ŭ*) and *ē* (*ĭ*) became diphthongs *ou* and *ei* in Northern French at an early period (after the 5th but before the 9th century); the 12th-century phonetic results *eu* and *oi* provided the present-day spellings, though the sounds thus represented have changed considerably since (compare *fleur* 'flower,' from *flour,* from *flore*). The greater

Table 10: Occurrence of Diphthongs Replacing Stressed Short Vowels in Romance Languages

Latin	*pĕde* 'foot'	*hĕrba* 'grass, herb'	*mŏrit* 'he dies'	*mŏrtem* 'death'
Sardinian	*pe*	*erva*	*mṓridi*	*morte*
Portuguese	*pe*	*herva*	*morre*	*morte*
Catalan	*peu*	*herba*	*mor*	*mort*
Occitan	*pe*	*erba*	*mor*	*mort*
French	*pied*	*herbe*	*meurt* (Old French *muert*)	*mort*
Italian	*piede*	*erba*	*muore*	*morte*
Romanian	—	*iarbă*	*moare*	*moarte*
Spanish	*pie*	*hierba*	*muere*	*muerte*
Rhaetian (Sursilvan)	*pei*	*jarva*	*miera*	*mort*
Rhaetian (Friulian)	*pid*	—	*mûr*	*muart*
Vegliot	*pi*	*jarba, jerba*	—	*muart*

extension of spontaneous diphthong formation in French than in other Romance languages is often attributed to the effects of the heavy stress presumably used by the Frankish superstratum.

Effects of nasal consonants on vowels

In nearly all Romance languages a following nasal consonant has caused peculiar development in a preceding vowel. In most cases the effect is limited to a raising or closing influence, but, in both French and Portuguese, phonological nasalization has taken place (*i.e.*, in both, a series of vowels distinguished by the presence of nasal resonance has developed). There, as well as in some other dialects (especially Chilean, Caribbean, and Andalusian Spanish, in the Romanian spoken in Albania, and in northern Occitan), nasal vowels are distinct from their oral counterparts and not mere variants (*i.e.*, they are phonemic). Thus, they serve to differentiate one meaningful form from another: *e.g.*, French *pin* 'pine,' pronounced /pɛ̃/ (ɛ stands for a short *e* sound, and the tilde sign [~] marks nasalization) versus *paix* 'peace,' pronounced /pɛ/; Portuguese *lã* 'wool' versus *la* 'there.'

Nasalization in both French and Portuguese was probably noticeable by the 10th century, though it may not have become phonemic until much later. Some claim that even today nasal vowel resonance is merely a surface manifestation of a latent underlying nasal consonant. It would appear that in both languages nasal vowels were more frequent in the Middle Ages than today; in roughly the 16th century in France, denasalization took place when the nasal consonant was intervocalic, and the *n* sound was retained—in, for example, French *bon* 'good (masculine)' (pronounced /bõ/) and *bonne* 'good (feminine)' (pronounced /bon/). In Portuguese the consonant did not always reappear after denasalization (compare *boa* 'good [feminine],' from *bõa,* from *bona*), though between *i* and *a* or *o* the palatal nasal consonant (pronounced somewhat like the *ny* in English *canyon*) is inserted (*vinho* 'wine,' from *vĩo,* from *vinu*).

Nasalization has sometimes, though without much conviction, been attributed to Celtic substratum influence. A better case can be made for the effect of such influence in the French *u* sound, /y/, pronounced like German *ü* or Greek upsilon [ʋ], though ignorance of Gaulish and certain chronological and geographic discrepancies make it difficult to argue in detail. The French /y/ is also found in most Occitan dialects (in which it may be a recent introduction from French), in Rhaetian, and in parts of Portugal and Italy; elsewhere it is sometimes a characteristic of affected speech.

Consonants. Another French pronunciation that is often imitated by socially pretentious speakers is that of the Parisian uvular *r* /ʀ/ (produced by vibration of the uvula, an appendage at the back of the mouth), which was not accepted in standard French until after the Revolution of 1789, though it was probably used by the Parisian bourgeoisie from the 17th century. It probably developed from the Latin double *-rr-*, differentiated from single *-r-*, which in Middle French tended to be pronounced with local friction. In most modern dialects of Provence the distinction between the two *r* sounds is still made (though Occitan dialects in general are adopting the French pronunciation). Brazilian Portuguese uses a similar contrasting pair of *r* sounds, with the usual trilled *r* represented in orthography by a single *r* and a velar, or "rough," *r* represented by *rr*: Brazilian *caro* 'dear' and *carro* 'cart.' Elsewhere only Puerto Rican Spanish and a few North Italian and Romanian dialects use the velar *r* regularly, though it is heard sporadically nearly everywhere.

Brazilian Portuguese and *r* sounds

One phonological development that is thought by many to be indicative of a very early split between the Eastern and Western Romance areas concerns the treatment of consonants between vowels. To the north and west of a line drawn between La Spezia and Rimini, in Italy, most dialects voiced Latin voiceless consonants between vowels and simplified geminates (doubled consonants); southern and eastern dialects to a greater extent retain the Latin voiced–voiceless–geminate system. The dividing line appears also to run through Sardinia, so that northern dialects are "Western" and southern ones "Eastern." Table 11 shows the treatment of intervocalic *p* and *t*.

Table 11: Development of Latin Intervocalic *p* and *t* in Romance Languages

Latin	*ripa* 'bank'	*rota* 'wheel'
Vegliot	*raipa*	—
Romanian	*rîpă*	*roată*
Italian	(Old Italian *ripa*)	*ruota*
Logudorian	*riba*	*roda*
Occitan	*riba*	*roda*
Catalan	*riba*	*roda*
Spanish	*riba*	*rueda*
Portuguese	*riba*	*roda*
French	*rive*	*roue* (Old French *ruede*)
Rhaetian		
Sursilvan	*riva*	*roda*
Engadine	*riva*	*rouda*
Friulian	*rive*	*ruede* (*rovede*)

Some believe that the voicing of voiceless sounds is connected with a similar, though not identical, process in Celtic known as lenition. Lengthening and subsequent development into diphthongs of accented vowels may be linked to the reduction of Latin doubled consonants to single consonants, as some theories suggest.

One noticeable difference between Latin and all the Romance languages is that the consonantal systems of the latter include a number of palatal and palato-alveolar consonants which did not exist in Latin. (Palatal consonants are formed with the tongue touching the hard palate; palato-alveolar sounds are made with the tongue touching the region of the alveolar ridge or the palate.) One consequence of the strengthening of the stress accent in the later Latin period was that unstressed *ĭ* and *ĕ* following consonants and followed by vowels became shortened to a nonsyllabic palatal *y* sound (called *jod*). The effects of this new sound on preceding consonants are varied, but in many cases these have been pronounced with the tongue raised more toward or against the roof of the mouth, or palate (a process classified as a form of assimilation), sometimes ending up eventually as a dental fricative (such as *z* and *th*) or affricate (such as *ch*) and perhaps modifying the preceding vowel. That this process began early is suggested by the not-infrequent confusion of *-tĭ-* and *-cĭ-* in orthography, sometimes represented even as *tz* in inscriptions. This palatal shift in pronunciation led to developments such as French *rouge,* Portuguese *ruivo,* Catalan *roig,* and Old Italian *robbio* from Latin *rubeum* 'red' and French *feuille,* Portuguese *folha,* Italian *foglia,* and Sardinian *fodza* from Latin *folia* 'leaf.'

Shifts involving palatal consonants

Another source of palatal consonants in Romance has been back (velar) consonants when immediately followed by a front sound: the velar consonant has often moved forward in the mouth, sometimes eventually to dental or alveolar position but often settling on a palatal or palato-alveolar position. This process, too, probably began early, first affecting velar consonants /k/ and /g/ preceding front vowels /e/ and /i/. That it had not occurred at the Classical period is shown by its absence in early loanwords into other languages (Berber, Basque, Celtic, Germanic, Albanian, and Greek). As central Sardinian dialects retain velar pronunciation in the environment of front vowels, it may be assumed that palatalization postdated the separation of the island from the rest of the empire. Vegliot evidence is difficult to interpret, as *ē* does not seem to have provoked palatalization, whereas *ĕ, ĭ,* and *u* did so. It was this sound change that resulted in the pronunciation of "soft" *c* before *e* and *i* (in most Romance languages this is an /s/ or /ts/ sound; in Italian and Rhaetian it is a /ch/ sound). Before *a, o,* and *u* the *c* retained its "hard" pronunciation (that is, a /k/ sound). In Classical Latin, before the sound change occurred, all *c* sounds were hard. Hence, Latin *centum* (/kentum/) gave rise to Italian *cento* (/chento/), Portuguese *cento* (/sento/), and Spanish *ciento* (/siento/ or, in Castilian, /thiento/).

In north-central France, Latin *a* must have advanced to a front position, with the result that it, too, palatalized preceding /k/ and /g/ sounds. The results give the palato-alveolar sounds of /sh/ and /zh/ (written in the International Phonetic Alphabet as /ʃ/ and /ʒ/, respectively), via /tʃ/ (as in English *church*), and /dʒ/ (as in English *jam*); *e.g.,* French *chanter* 'to sing' developed from Latin *cantare, joie* 'joy' from *gaudia.* West Rhaetian dialects show a similar development (compare Sursilvan *tgaun,* Engadine *chaun,* French *chien,* from Latin *canem* 'dog'), as do Franco-Provençal and Northern Occitan dialects, but Picard and some Norman dialects do not (Picard *canter,* with an unpalatalized *c,* from Latin *cantare*; *kier* 'dear,' from *carum*). The change is assumed to have taken place at a later period than the palatalization of *k* when followed by *e* or *i,* which did not affect Frankish words. These, on the other hand, succumbed to the type of palatalization in which *k* /k/ changed to *ch* /tʃ/ and then to *sh* /ʃ/ (**skina* becomes *échine* 'backbone').

In Romanian, velar consonants were moved forward under the influence of a following *i* and *e,* and dental consonants were moved back to a palatal position under the same influence—*e.g., ţară* from *terram* 'earth'; *şi* 'and' from *sic* 'thus'; *cer* from *caelum* 'sky.' Labial consonants are also affected in some dialects: *k'ept* from *piept* from *pectum* 'chest'; *jin* from *vin* from *vinum* 'wine.' Romanian also has, in final position, a series of "soft" consonants. These are transparently derived from earlier "hard" consonants followed by *i,* performing certain important morphological functions: *lupi* /lup'/ 'wolves' from *lup* /lup/ 'wolf'; *cînţi* /kɨnts'/ 'you sing' from *cînt* /kɨnt/ 'I sing.'

Palatalization of consonants in Romance was effected not only by following front vowels but also by juxtaposed front consonants, especially when a velar (such as Latin /k/ or /g/) was next to a dental (such as /t s n/) or a lateral (/l/) in medial position, sometimes as a consequence of the loss of an unaccented vowel during the Vulgar Latin period. Results of this process vary from language to language. Table 12 gives examples of these changes.

Juxtaposed front consonants

Table 12: Results of Palatalization of Consonant Clusters

Latin	*noctem* 'night'	*coxam* 'hip'	*piscem* 'fish'	*pugnum* 'fist'	*oc'lum* 'eye'
Vegliot	*nwat*	—	*pask*	—	*vaklu*
Romanian	*noapte*	*coapsă*	*peşte*	*pumn*	*ochi*
Sardinian	*notte*	*koša*	*piske*	*pundzu*	*okru*
Italian	*notte*	*coscia*	*pesce*	*pugno*	*occhio*
Occitan	*nôit, nuech*	*cuoissa*	*peis*	*ponh*	*uelh*
Catalan	*nit*	*cuixa*	*peix*	*puny*	*ull*
Spanish	*noche*	*cojo*	*pez*	*puño*	*ojo*
Portuguese	*noite*	*coxa*	*peixe*	*punho*	*olho*
Rhaetian					
Sursilvan	*notg*	*queissa*	*pesch*	*pugn*	*egl*
Engadine	*not*	*cossa*	*pesch*	*puogn*	*ögl*
Friulian	*ñot*	*cuesse*	*peš*	*pugn*	*voli*
French	*nuit*	*cuisse*	(*poisson*)	*poing*	*oeil*

It will be noted that in Romanian a labial consonant has been substituted for the velar in the Latin clusters *-ct-*, *-x-* /ks/, and *-gn-*: *piept* from *peptum* 'chest,' *coapsă* from *coxa* 'thigh,' *lemn* from *lignum* 'wood.' Perhaps there was first assimilation of the velar to the dental—as in Italian *-tt-* from Latin *-ct-* and Sardinian *-nn-* from Latin *-gn-* (*linna* from *ligna* 'line')—followed by differentiation of the first element of the geminate. It is notable that Latin *l* regularly becomes *jod* /y/ after another consonant in Italian (*piacere* from *placere* 'to please'; *fiore* from *flore* 'flower'; *chiave* from *clave* 'key'; *ghianda* from *glanda* 'acorn') and after velars in Romanian (*plăcea, floare,* but *cheie* /kjej/, *ghindă* /gjində/). In Spanish and Portuguese a following *l* in Latin often palatalizes labial consonants (*p, f*) as well as velars, in initial as well as medial position; *e.g.,* Latin *planum* becomes Spanish *llano* 'plain,' Portuguese *chão;* Latin *afflare* becomes Spanish *hallar* 'to find,' Portuguese *achar.*

Grammar. Item for item, the Romance languages all appear grammatically close to Latin and to each other: superficial resemblances in individual expressions may, however, mask differences of content and construction that are difficult to describe. The most obvious difference between Latin and Romance is in the comparative autonomy of morphemic units, especially words. In Romance, Latin inflectional endings have been much reduced, and more reliance is placed on syntactic construction to convey sentence meaning; that is, Romance languages are more "analytic" than the predominantly "synthetic" Latin. A corollary of this is that word order is less flexible in Romance, as it has become the principal means of showing relationship between words in the sentence.

Gender marking

The reduction of inflectional endings. The inflectional endings have been lost most in nouns and adjectives. The Classical Latin five-case declensional system has everywhere been replaced (with a couple of doubtful exceptions) by a two-gender system, in which normally masculine gender is marked by survivors of the second (*-us*) declension endings of Latin (Italian *cavallo,* Portuguese *cavalu,* Romanian *cal,* Sardinian *kaḍḍu,* Rhaetian *cavagl,* from Latin *caballus* 'horse'), and feminine is marked by first (*-a*) declension endings (Italian *capra,* Spanish *cabra,* Rhaetian *caura,* Romanian *capră,* from Latin *capra* 'goat'). Cognates of third-declension Latin noun forms are incorporated into the same system, but their gender is marked by changes in the article or accompanying adjective (agreement or accord) rather than by overt markers in the word itself (for example, masculine Italian *il monte,* Catalan *el munt,* from Latin *mons, montem* 'mountain'; feminine Italian *la notte,* Catalan *la nit,* from Latin *nox, noctem* 'night'). In modern French, although gender is marked in the written language, however inconsistently, by the presence or absence of final *-e,* any overt morphological markers the spoken language may have are more complex in character, and more reliance is placed on syntactic agreement; thus, *chatte* 'she-cat' is distinguished from *chat* 'cat' by the presence or absence of the final consonant sound *-t* in pronunciation, but *(le) tour* 'tour, trick' and *(la) tour* 'tower' have identical phonetic shapes though they belong to different gender classes.

All the Romance languages continue to mark plurality in nouns and adjectives morphologically, though in modern spoken French this is not done consistently. In Western Romance the sign of the plural is usually *-s,* derived from the Latin accusative plural inflection: Spanish *caballos, cabras, montes;* Occitan *cavals, cabras, mons;* Catalan *cavalls, cabres, munts;* Sardinian *kaḍḍos, krabas, montes;* Old French *chevals, chèvres, monts.* In Italian and Romanian, however, plurality is shown by a final *-i* (which in Romanian "softens" the preceding consonant) or, in the case of some feminine nouns, by a final *-e:* Romanian *cai, capre, munți, nopți;* Italian *cavalli, capre, monti, notti.* These endings may derive from Latin nominative plural first- and second-declension endings *-ae* and *-ī,* or they may represent a somewhat irregular development of the *-s,* favoured elsewhere.

The loss of the case system. The Latin nominal case system has disappeared in all modern languages except Romanian, in which the inflected article distinguishes the nominative and accusative from the genitive and dative (see Table 13). Thus, when other Romance languages would use a preposition to indicate a certain relationship between words, Romanian resembles Latin in using an inflected form (*e.g.,* Latin *matris* 'the mother's' becomes Romanian *mamei,* French *de la mère,* Italian *della madre*).

Table 13: Declensional System of Romanian

	singular	plural
Masculine 'son'		
Nominative-accusative	*un fiu, fiul*	*fii, fiii*
Genitive-dative	*unui fiu, fiului*	*unor fii, fiilor*
Feminine 'mother'		
Nominative-accusative	*o mamă, mama*	*mame, mamele*
Genitive-dative	*unei mame, mamei*	*unor mame, mamelor*

The role of word order

In Old French and Old Provençal some remnants of a case system remained, in that the masculine nominative (subject of the verb) was distinguished from the other cases (collectively called oblique). Such grammatical information is conveyed by word order in most modern Romance languages, as in English, with the subject normally preceding the verb: French *Pierre appelle Paul* 'Peter calls Paul'; Portuguese *Pedro chama Paulo;* Italian *Piero chiama Paulo.* Some Romance languages pick out the object of the verb, if it is a person, by an additional particle: Spanish *Pedro llama a Pablo;* Romanian *Petru cheamă pe Pavel.* Several Italian dialects, as well as Sardinian and occasionally Engadine and Portuguese dialects, have similar constructions: Calabrian *Chiamu a Petru* 'I call Peter'; Elba *Ò visto a ttuo babbo* 'I saw your grandpa'; Engadine *Amè a vos inimihs* 'Love your enemies.' It is notable that the Italian-based lingua franca used by Mediterranean sailors since the 16th century also picks out the personal object (*e.g., Mi mirato per ti* 'I saw you').

The emergence of articles. The definite and indefinite articles were unknown in Latin but developed everywhere in Romance, usually from the Latin demonstrative *ille* 'that' (though in a few parts from reflexive *ipse* 'himself') and the numeral *unus* 'one.' The definite article is proclitic (attaches to the following word) in most Romance languages (*e.g.,* Italian *il monte*); in Romanian it is enclitic (*e.g., muntele* 'the mountain'). The articles seem to have played some part, during the older stages of the languages, in distinguishing subject from object; the article is more often used where a Latin nominative would have occurred than in other cases, perhaps to give prominence to the topic of the sentence. Today the use of the article has so extended that such distinction is no longer possible; in French, for instance, a common noun is always accompanied by a determiner such as an article, demonstrative, or possessive, so that forms remaining from the earlier stage, such as *avoir faim* 'to be hungry' (literally, 'to have hunger'), are often regarded as idiomatic and inexplicable in terms of modern structure.

The survival of verbal inflection. In the passage from Latin to Romance, verbal inflection has survived much more than noun declension. Although the four regular Latin conjugations have been virtually reduced to two, with only the *-a-* class remaining truly productive, other features of the verb seem almost unchanged. In most languages, for instance, the person markers are directly traceable to Latin origins (*i.e.,* to Latin *-ō, -s, -t, -mus, -tis, -nt*). Modern spoken French is the only major language in which the personal endings no longer serve the same function as in Latin. Today, person is marked in French principally by pronouns derived mainly from the Latin emphatic nominative forms of the personal pronoun: *J'aime* /ʒɛm/ 'I love,' *tu aimes* /tyɛm/ 'you love' from (*ego*) *amo,* (*tu*) *amas.* The creoles have taken this process even further, in that their verb forms are usually invariable but are prefixed by elements indicating person, tense, aspect, and so on, as in many West African languages: Louisiana French /motegẽ/ 'I was having' from *mon* /mo/ *étais* /te/ *gagner* /gẽ/; and similarly /ilagẽ/ 'he will have.'

Verb conjugations

In the metropolitan languages, verbal modalities are shown, as in Latin, by inflection. Some Latin verb endings, such as that of the *-r* passive or of the future, have disappeared; others, such as the pluperfect indicative and subjunctive, have survived in a few languages with modified function. But most modern languages have reflexes of the present, perfect, and imperfect indicatives and of one or more subjunctive tenses. The imperfect indicative, a Latin innovation, survives almost intact, though the evolution of its form, not to mention its function, presents problems. The *-ī-* stem form in Latin *-iēba-* is thought to have coalesced early with the *-ē-* stem *-ēba-* form, but a few modern languages (notably Italian, Friulian, and some Spanish and Portuguese dialects) have reflexes of an *-ība-* form that might have survived from popular Latin. The Latin *-āba-* form survives almost everywhere, though in most French dialects its older reflexes, *-eve* and *-oue,* have been replaced in modern times by forms derived from Latin *-ēba-.* These latter are thought to be widespread but are puzzling phonologically as they have very often irregularly lost their *-b-* (Spanish, Portuguese, and others *-ía,* French *-ais*).

The Latin perfect of the type *amāvit* 'he has loved' is known by all the literary languages but is rare in speech in French, Italian, and Romanian, in which it has been replaced by a new compound past made up of the verb for 'to have' and a past participle. The latter structure is known to some extent in all Romance languages, often being used to express a more recent past than the preterite *amāvit* form, which also indicates action in the past (without reference to duration or repetition): Romanian *am cîntat,* Italian *ho cantato,* French *j'ai chanté,* Spanish *he cantado,* Old Portuguese *hei cantado,* Engadine *ha chantà, hè chantò,* Sardinian *kantau appo,* from Latin *habeo cantatum* 'I have sung.' In modern Portuguese the preferred auxiliary is *ter* 'to have, to hold' rather than *haver,* producing forms such as *tenho cantado,* while modern Catalan has two forms of the perfect, the pan-Romance type (*he cantat*) and a specific type that uses the verb for 'to go' plus the infinitive (*vaig cantar*), semantically different.

The disappearance of the Latin future has been remedied in most Romance languages by the development of new forms of periphrastic origin. Many of these forms use some reflex of *habēre* 'to have' joined to an infinitive. From Latin *cantāre habēo* 'I will sing' are derived Italian *canterò,* Spanish, Catalan *cantaré,* Portuguese *cantarei,* French *je chanterai,* Rhaetian *c(h)antero, c(h)antera,* Occitan *cantarai*; *habēo cantāre* gives southern Italian *aggio cantà* (similar forms are seen in earlier Spanish, Portuguese, and northern Italian). Latin *habēo ad cantāre* produces Sardinian *ap' a kantare,* and *habēo de cantāre* gives Portuguese *hei-de cantar* (more popular than *cantarei*).

A periphrastic future of the type shown in English 'I'm going to sing' enjoys popularity in Romance, mainly to indicate a less distant future event than the more formal future tense (*e.g.,* French *je vais chanter,* Spanish *voy a cantar*). Other periphrases used in Romance are 'I will (wish to) sing,' as in Romanian *voi cînta*; 'I must sing,' as in Sardinian *deppo kantare*; 'I'm coming to sing,' Sursilvan *jeu vegnel a cantar*; and 'I have that I should sing,' as in popular Romanian *am să cînt.* Notably, Dalmatian does not seem to know periphrastic Romance futures but uses a form *kantuora* (perhaps from Latin *cantāverō*) as both future and conditional.

The conditional form

The Romance conditional, or "future in the past," a form not found in Latin, is in many languages related to the new future tense. In the Western languages it is composed of the future stem (or infinitive) plus a past-tense marker related to reflexes of *habēre.* In some cases an imperfect form is used, in others a perfect form; examples are French *je chanterais* 'I would sing,' Spanish, Portuguese, Occitan, and Catalan *cantaría,* and Italian *canterei, -ebbe,* and so on. In Romanian the conditional marker can either precede or follow the infinitive and may be derived from the imperfect of *vrea* 'to wish': for example, *aş cînta, ar cînta,* and so on, or (obsolete and dialectal) *cîntare-aş, cîntare-ar,* and so on.

Syntax. Word order is the means most used by modern Romance languages to show the grammatical relationship between words; statistically the most frequent order in statements is subject–verb–object. In many of the Romance languages, interrogation can be shown by inversion of the subject and verb, placing the verb, as the element on which the interrogation falls, at the beginning of the sentence (Spanish *¿Vino el hombre?,* Italian *É venuto l'uomo?* 'Has the man come?'). In such examples, however, it is the intonation (represented in writing by the question mark) rather than the word order alone that marks the question. Inversion, without interrogative intonation, is not infrequent in emphatic assertions. Unambiguous question markers—such as the Latin particles *-ne, nonne,* and *num*—are lacking in most Romance standards; popular speech, though relying everywhere principally upon intonation, often has developed new particles to reinforce interrogation. Romanian has *oare* (*Oare a venit?* 'Has he come?'); Italian uses dialectal *ce, che,* or *o* (Vulgar Tuscan *Che è venuto?* 'Has he come?'; *O come si chiame?* 'What is he called?'); Sardinian has *a* (*A morde kkŭstu kăne?* 'Does this dog bite?'); and French and Limousin have *ti* (generalized from such forms as *a-t-il?;* French *Je suis-ti bête?* Limousin *Sieu-ti nesci?* 'Am I stupid?'). In modern standard French great use is made of *est-ce que* as an interrogative particle: *Est-ce qu'il est venu?* 'Has he come?' *Comment est-ce qu'il s'appelle?* 'What is his name?'

Negation in Latin was expressed by a range of special items (*non, nemo, nihil, nullus, nunquam,* and so on). Although some of the others survive in Romance, continuators of *non* are usually used for negative expression and are regularly prefixed to the verb. Nuances within negation are usually expressed by the adjunction of other items. In France, both north and south, and in northern Italy and some of the Swiss Rhaetian areas, the *non* particle has been so weakened phonetically that it no longer can express unambiguously the important distinction between negative and positive; hence, formerly positive adjuncts have acquired its negative meaning.

French *personne / une personne* signifies 'no one / a person'; *pas / un pas* means 'not / a step'; and *plus* can mean 'more / no more.' In popular speech the *non* particle is frequently omitted altogether in areas that use these additional forms (*e.g.,* French *Je (ne) le vois pas*; Occitan *Lou vese pas* for *Noun lou vese* 'I don't see it').

The reduction of the subjunctive. Morphologically, the verb system survived comparatively intact from Latin to Romance; if the schoolbooks, heavily influenced by Latin grammar, are right, the ways in which the verb forms are used are not so very different from Latin either. The most obvious change has been the reduction of uses as well as of forms of the subjunctive, with, at the extreme, modern French treating them as automatically determined variants to be used obligatorily after certain phrases and conjunctions and virtually eliminating tense differences within the subjunctive mood. When the subjunctive retains a function in Romance—that is, in contexts in which it can contrast with the indicative—it has developed emotive overtones, especially suggesting doubt, unreality, or some sort of hypothetical futurity. It is used especially in subordinate clauses dependent on verbal expressions of command and exhortation, emotion, or doubt: Romanian *Vreau să vină* 'I want him to come'; Engadine *Mieu bap voul ch'eau lavura* 'My father wants me to work'; French *Je doute qu'il vienne* 'I doubt that he's coming'; Portuguese *Duvido que seja feliz* 'I doubt that he is happy'; Italian *Temo che sia tarde* 'I'm afraid it's late'; Spanish *Temo que él lo diga* 'I'm afraid he'll say it.' The subjunctive also regularly follows subordinating conjunctions that project action forward into the future, notions such as 'until,' 'before,' 'in order that': French *avant que vous soyez venu* 'before you came'; Spanish *hasta que sea feliz* 'until he is happy'; Italian *perchè potessi fare in tempo* 'so that I might do it in time'; Portuguese *antes que eu o veja* 'before I see it'; Catalan *abans que vingui* 'before he comes.'

Modern French and the subjunctive

On the whole, however, the Romance languages use the subjunctive less frequently than does Latin, with recession particularly, when no doubt is implied, in indirect speech and in temporal and concessive clauses (in French, use of the subjunctive after concessive conjunctions such as *bien que* and *quoique* 'although' was imposed by 18th-century grammarians). The infinitive is often used in subordinate constructions where Latin would have used a subjunctive—*e.g.,* French *dites-lui de s'en aller,* for *dites-lui qu'il s'en aille* 'tell him to go away.' Romanian, on the other hand, has even extended the use of the subjunctive in such constructions, perhaps reflecting a substratum influence that is felt, too, in some Balkan languages. Greek influence is sometimes credited with similar constructions (usually using the indicative rather than the subjunctive) found in northeastern Sicily, northern Calabria, and the Salentine Peninsula.

Conditional clauses. One area of syntax in which the Romance languages vary widely in the extent to which they retain and in the manner in which they replace the Latin subjunctive is that of past-tense hypothetical conditional clauses. The Latin formula *si habuissem dedissem* 'if I had had it, I would have given it,' though challenged by a type using the indicative tense since Ciceronian times, has sporadically survived into Romance, especially in the older stages of the languages and in scattered languages of southern Italy (*Se potessi, facessi* 'If I could, I would do it'), Rhaetian (Sursilvan *Jeu vegness, sche jeu vess peda* 'I'd come, if I had time'), and Romanian (*dacă aş fi avut destui bani, aş cumpărat-o* 'If I had had enough money, I would have bought it').

In most languages, however, a new conditional form replaces the subjunctive in "if" clauses. Thus, in Spanish, Portuguese, and most Italian dialects, sentences of this type are seen: Spanish *si yo tuviese bastante dinero, lo compraría*; Italian *se avessi abbastanza denaro, lo compre-*

rei; Portuguese *se tivesse bastante dinheiro compraríao* ('if I had enough money, I'd buy it'). Spoken Catalan usually prefers a similar construction (*si estudiessis ho sabries* 'if you studied, you would know it'). Another construction that replaces the subjunctive by the imperfect indicative in the "if" clause is normal in Catalan and in French as well as in Corsica and Sardinia: Catalan *si estudiaves ho sabries* ('if you studied, you would know'); French *si j'avais assez d'argent, je l'achèterais* ('if I had enough money, I'd buy it'); Logudorian *si denía abba deo dia buffare* ('if I had water, I'd drink'). Other constructions using the imperfect indicative or the conditional in both clauses are found mainly in substandard styles—both types are common in French and Romanian, the former in Tuscany, southeastern Italy, and Spain and the latter in much of southern Italy.

The imperfect indicative

Morphology. Romance methods of forming new words from native sources are in part inherited from Latin (the morphological device of adding a suffix and that of prefixing an element that modifies the original meaning) and in part later developments (mainly that of combining two or more free forms to make compound words and of changing or extending the syntactic distribution of an already existing word).

Derivation by means of suffixes is the most popular and widespread device; verbs in particular must be morphologically marked as members of a conjugation, of which those corresponding to Latin *-āre* form by far the most frequent and indeed in modern times virtually the only productive class (thus Latin *plantāre* 'to plant,' Italian *plantare*, Engadine *plaunter*, French *planter*, Catalan *plantar*, from *planta* 'plant'). Infixes, inserted between the verbal root and the conjugation marker, are common. Sometimes they continue Latin infixes, such as the frequentative (compare *jactāre* for *jacere* 'to throw,' Italian *gettare*, French *jeter*, Catalan *gitar*, etc.); sometimes they add semantically to the root meaning (compare pejorative Italian *lavoracchiare* 'to slack off' from *lavorare* 'to work,' French *criailler* 'to bawl' from *crier* 'to cry'). The Greek verbal infix *-iz* (as in English suffix *-ize*) is particularly popular in modern Romance languages (*e.g.*, *automatiser*).

Among noun suffixes, diminutives are frequent and, except perhaps in French, still productive. Romanian uses *-aş* (*baieţaş* 'little boy'), but the other languages prefer derivatives of Latin *-ittus* (especially in Spanish: *arbolito* 'little tree,' *señorita* 'Miss, young lady,' etc.; but also French *sachet* 'little sack,' Italian *foglietta* 'little leaf,' etc.) or of Latin *-īnus* (preferred in Italian: *tavolino* 'little table, desk,' *signorina* 'young lady'; and Portuguese: *copinho* 'little drinking glass,' *senhorinha* 'young lady'). The Latin *-ōne* suffix has, conversely, acquired augmentative meaning in several languages (Italian *cavallone*, Spanish *caballón* 'large horse'). Romanian *-oi, -oaie* seems to continue an adjectival form of this suffix *-oneus, -a*; *căloi* 'large horse,' *căsoaie* 'big house.'

Other frequent suffixes sometimes have a "learned" modern form alongside the older "popular" one—*e.g.*, Latin *-atione* becomes Italian *-agione / -azione*, French *-aison / -ation*, Spanish *-azón / -ación*, Portuguese *-azão / -ação;* also Romanian *-ăciune / -aţiune* (*-aţie*), and Occitan *-azó*.

Suffixes that remain extremely productive include the Latin verbal adjectival *-bilis* (not found in Romanian), which can be seen in Italian *bastevole* 'enough,' French *admirable*, Spanish *amable* 'pleasing'; and Latin verbal nominal *-mentum*, which can be seen in French *abonnement* 'subscription,' Spanish *cobijamiento* 'lodging,' Italian *abboccamento* 'interview, parley,' Romanian *acoperămînt* 'cover.'

Prefixes

Prefixing of modifying elements remains frequent in all languages (Italian *autostrada* 'highway,' Spanish *contraveneno* 'antidote,' French *photocopie* 'photocopy'), although some older prefixes may hardly be recognized as such today. The "repetitive" verbal prefix *re-* remains particularly active (Romanian *răpune* 'to kill,' Italian *ricattare* 'to recover,' French *racheter* 'to buy back').

Compound words, though less frequent than in the Germanic languages, are not uncommon (*e.g.*, French *cheflieu* 'principal town,' Italian *primavera* and Romanian *primăvară* 'spring,' Spanish *lavamanos* 'wash basin').

Originally a compounding process, the most common method of forming adverbs from adjectives (suffixing of Latin *mente* 'mind') has become in most languages a morphological process, although Spanish and Portuguese retain traces of the earlier stage in phrases such as *severa e* (*y*) *cruelmente* 'severely and cruelly.'

Among the syntactic means that most Romance languages use to extend vocabulary is the potent device, unavailable to Latin, of juxtaposing to any part of speech an article or other determiner and using it as a noun (*e.g.*, Italian *il perchè* 'the reason,' Spanish *lo útil* 'utility, something useful,' French *un je ne sais quoi* 'an I-don't-know-what'). In French and Spanish, verbal infinitives are frequently so treated (*le devoir* 'duty,' *el poder* 'power,' etc.); Romanian also uses infinitives as verbal nouns, but they are differentiated formally by retaining the full form (*e.g.*, *cîntare* 'singing'), compared with the shortened verbal form (*cînta*). In earlier stages of most Romance languages the verbal root (most often as it appears in the third-person singular present indicative) could be used as a noun, a process known as back-formation (compare Romanian *laudă* 'praise,' Italian *domanda* 'question,' French *approche* 'approach,' *désir* 'desire,' Spanish *baila* 'dance,' Portuguese *muda* 'change').

Just as former adjectival forms are frequently used as substantives, so are nouns used with adjectival function; there seem to be few restrictions on this use, though practice varies as to whether agreement should be made (French *les frères ennemis* 'enemy brothers,' with agreement; *une femme médecin* 'a woman doctor,' without agreement). Past-participial forms normally act as adjectives, as in English.

Romance makes use of gender classification to extend and modify its vocabulary, especially by relating the gender markers to sex differences (*e.g.*, Romanian *nepot, nepoată*, Occitan *nebut, nebudo*, Spanish *nieto, nieta*, Portuguese *neto, neta*, Catalan *net, neta* 'nephew, niece,' with Italian invariable *nipote*, and French lexically differentiated *neveu, nièce*). Modern French makes particularly fruitful use of gender differences (originally via ellipse); thus, *le* (*vin de*) *champagne* (the drink) / *la Champagne* (the province); *La Normandie* (the province) / *le Normandie* (the ship).

Vocabulary. *Latin inheritance.* The basic vocabularies (the most frequently used lexical items) of all the Romance languages are in the main directly inherited from Latin. This applies equally to "function" words, such as *de* 'of, from' (Romanian *de*, Italian *di*, Rhaetian *da*, French *de*, Spanish *de*, Portuguese *de*), as to common lexical items, such as *facere* 'to do' or *aqua* 'water' (Romanian *a face, apă*, Italian *fare, acqua*, Logudorian *fágere, abba*, Engadine *fer, ova*, French *faire, eau*, Catalan *fer, aigua*, Spanish *hacer, agua*, Portuguese *fazer, água*). In some cases different Romance languages inherit words perhaps from different strata of Roman society. Thus, for 'lamb,' forms derived from Latin *agnus* remain in southern Italy and Galician (*año*), but forms derived from diminutive *agnellus* prevail in Romanian (*miel*), Italian (*agnello*), French (*agneau*), Rhaetian (Engadine *agnê*, Friulian *añel*), Occitan (*anhel*), and Catalan (*anyell*), with Sardinian and some Calabrian dialects using another form derived from Latin *agnone* (such as Logudorian *andzone*). Spanish and Portuguese, however, prefer a derivative of a different word, *chorda* (*cordero, cordeiro*), referring perhaps to the birth process; this word is also found in Occitan and Catalan.

Varying inheritance

Pre-Romance borrowings. Some words shared by most of the Romance languages are not of Latin origin but were probably borrowed from other languages before Latin unity was disrupted, especially words of Celtic origin, such as Latin *carrum* 'cart,' Romanian *car*, Italian *carro*, Logudorian *karru*, Rhaetian *k'ar*, French *char*, Occitan and Catalan *car*, Spanish and Portuguese *carro*.

In Christian Latin a great many Greek ecclesiastical terms were borrowed, which survived in most Romance languages. For example, the Greek word *episkopos* (literally, 'overseer') was borrowed into Latin as *episcopus* 'bishop,' which gave rise to Vegliot *pasku*, Logudorian *pískamu*, Italian *vescovo*, Engadine *ovaisch*, Friulian *veskul*, French *evêque*, Occitan *avesque*, Catalan *bisbe*, Spanish *obispo*, and Portuguese *bispo*.

Germanic words did not penetrate into Latin very frequently before the separation of the various Romance languages from Latin, so that few of them have more than limited extension. Only one Germanic word is known for certain to be found in both Eastern and Western Romance—*sapōne* 'soap,' recorded in Pliny and occurring as Romanian *săpun,* Vegliot *sapaun,* Italian *sapone,* Logudorian *sabone,* Engadine *savun,* French *savon,* Occitan and Catalan *sabó,* Spanish *jabón,* and Portuguese *sabão.*

Later Latin borrowings. Many Latin words are widespread throughout the Romance languages even though they do not date back directly to the imperial period; these are the "learned" words that have freely entered the languages at virtually every period, borrowed from Latin used as a scholarly language. Because of this later borrowing, such words as *capital, natura, adulterium,* and *discipulus* appear in Romance virtually unchanged from Latin, as they do in other European languages; Romance Latinisms, however, are quite normally used in contexts in which similar words would sound stilted and pedantic in English (*e.g.,* French *supprimer* 'suppress' but often used to mean 'to do away with').

Vocabulary variations. However similar the Romance vocabularies are to each other, considerable differences nevertheless exist. Some of these may be traced to imperial times, when provinces may have developed their own vocabulary preferences. For instance, for 'oak' Eastern Romance seems to have preferred Latin *quercus* (Logudorian *kerbu,* southern Italian *quercia,* etc.), whereas the West preferred the alternative *robur* (Italian *rovere,* Occitan and Catalan *roure,* Spanish and Portuguese *roble,* Old French *rouvre*—modern French *chêne* is of Celtic origin, while Romanian *stejar* is perhaps of Balkan origin). In some cases the conservative peripheral areas have retained a word that was displaced in more central regions; thus, for 'beautiful,' *formosus* is preferred in Romanian (*frumos*), Spanish (*hermoso*), and Portuguese (*formoso*), whereas *bellus* is more popular in Vegliot (*bial*), Italian (*bello*), Rhaetian (*bal, biel*), French (*beau*), Occitan (*bel*), and Catalan (*bell*).

Borrowing from substratum and superstratum languages

When Romance borrowed vocabulary from the substratum, differentiation must have taken place early (certainly before the indigenous languages died out). Thus, Spanish *vega,* Portuguese *veiga* 'wooded ground by a river' (probably from a non-Indo-European Iberian language, compare Basque *ibaiko* 'riverbank'), French *charrue* 'plow,' *borne* 'boundary stone' from Celtic, and Romanian *barză* 'stork' (perhaps from Dacian, compare Albanian *bar*) probably were used during Roman times in some form. The debt of Romance vocabulary to substratum languages is probably not too great but is difficult to estimate with any certainty. When there is no known source form or cognate for a word, scholars often suggest an Iberian, Dacian, Ligurian, or Gaulish origin, but, as little is known of these languages, some such theories are mere speculation.

After the influx of barbarian invaders, Romance vocabularies differentiated further as each borrowed from its own superstratum (language superimposed upon Romance). French, for instance, is estimated to have taken some 700 words from Frankish (a Germanic language), not all of which have survived but some of which have passed via French into other Romance languages. Many of these were concerned with agriculture (*jardin* 'garden,' *houe* 'hoe,' *blé* 'wheat,' *gerbe* 'sheaf,' etc.) or with war (*guerre* 'war,' *heaume* 'helmet') or social organization (*sénéchal* 'seneschal,' *chambellan* 'chamberlain,' *maréchal* 'marshal,' *baron* 'baron'). The occupation of much of northern Italy by speakers of Langobardic (also a Germanic language) left less of a mark on Italian vocabulary, though dialects retain more words (estimated at some 300) than does the standard language. Standard Italian borrowed little in the way of administrative or military terms but accepted a number of words from rural life (*melma* 'mud,' *zecca* 'sheep tick,' *stamberga* 'hut,' etc.). The Visigoths, who occupied Iberia, were more Romanized than the other Germanic invaders and indeed had abandoned their Germanic tongue by the 7th century AD. Thus, borrowings from Visigothic into Spanish and Portuguese are less frequent, though still not inconsiderable; some (such as *estaca* 'stake,' *brotar* 'to bud') are common to all the languages of the Iberian Peninsula.

Slavic infiltration into the Balkans led Romanian to adopt a very large number of Slavic words, some in the basic part of the vocabulary. At exactly what stage in history they were borrowed is uncertain, for the earliest Romanian texts, of the 16th century AD, are saturated with Slavic terms from different dialectal sources, though South Slavic predominates. Possibly the borrowings occurred after the 9th century, when the Hunnish Bulgarians, who had adopted Slavic speech, established a powerful state and embraced Christianity, and Slavic pressures were already very strong. Among common Romanian words of Slavic origin one may mention *a trăi* 'to live,' *hrană* 'food,' *ceas* 'hour,' *bogat* 'rich,' *prieten* 'friend,' *a munci* 'to work.' The Magyars (modern Hungarians) also lent a smaller number of words to their Romanian neighbours (*e.g., oraş* 'town').

Words borrowed from Arabic

Islāmic invaders into Europe from the 8th century had considerable effect on the vocabulary of the Western Romance languages, even though occupation was confined to southern regions. With its superior cultural and agricultural skills, the Arab world had much to teach Europe of the early medieval period. Words entered via two routes, Sicily and Spain, and usually their form gives clues about their provenance—if the Arabic definite article (*al*) has coalesced with the root, the word is from Moorish Spain (thus Spanish *algodón* 'cotton,' Portuguese *algodão,* Old French *auqueton* via Spain, but Italian *cotone,* French *coton* via Sicily). The Arabs introduced into Europe many exotic plants and fruits and with them their names, such as oranges (Spanish *naranja*), lemons (Spanish *limón*), and artichokes (Spanish *alcachofa,* Italian *carciofo*). In some cases the Iberian Peninsula has adopted the Arabic word for such plants, while other languages prefer words of other origin—'rice' is *arroz* in Spanish and Portuguese, *arròs* in Catalan, but Italian and French prefer a Greek word (*riso, riz*), as do Vegliot (*rize*), Rhaetian (Friulian *ris*), and Romanian (*orez*). Apart from the numerous Arabic words known throughout Romance (especially words for 'algebra,' and the like), many are peculiar to the Hispanic languages, including such administrative terms as Spanish *alcalde* 'mayor' or *alguacil* 'senior police officer' and such commercial terms as *almacén* 'warehouse, department store.'

Many of the words individual languages borrowed from other sources or fashioned themselves from native sources did not remain private property for long. Interchange among the Western languages has been common since the earliest times and especially from the 16th century. Perhaps French has been the greatest supplier of words throughout the ages, often displacing native words. But French, too, has borrowed heavily from the other languages, especially when they have been purveyors of new objects (such as *patate, banane, tabac,* introduced into Europe by Spanish and Portuguese explorations) or of special cultural values (Italian musical and architectural terms, as well as words to do with banking). Borrowing into minor languages from prestigious neighbours has, naturally, been prolific. Passage of words in the other direction is rare and usually employed for comic or other emotive effect (though Occitan in its heyday supplied a good many words of all sorts—even, it is said, *amour* 'love' to French).

Borrowings from non-Romance languages are less frequent and often frowned on by purists but far from negligible. Any contact in specialized spheres has produced a crop of loanwords, especially since the 17th century, when French in particular began to borrow a fair number from its Germanic neighbours. In recent times, the influx of Anglicisms has become a flood, resisted to the death by some purists. Many of these, however, are ephemeral or specialized, and none affects the basic vocabulary in which Latin-inherited words continue to predominate.

Orthography. Today the Romance languages are all written in the Latin alphabet, with certain modifications, though until the mid-19th century Romanian was normally written in Cyrillic (used in Moldova until 1989), and, in the Middle Ages, Arabic script was used for some Spanish dialects.

As soon as scribes first made attempts to write in vernacular languages, they found the resources of the Latin alphabet inadequate to represent the non-Latin sounds of their spoken language. One device used to overcome these difficulties was the addition of the letter *h* to another, to indicate a deviant pronunciation: thus, *ch* might represent the /ch/ sound in Spanish (*e.g., muchacho* 'boy') or the /sh/ (earlier /ch/) sound in French (*e.g., chef* 'chief'). *C* would normally be used for the /k/ sound (before *a, o, u*) or an /s/ or /th/ sound (before *i, e*). In Italian, conversely, *ch* serves to distinguish the /k/ sound, followed by *e*, from the /ch/ sound (compare *che* /ke/ 'that, who' with *c'è* /che/ 'there is'). *H* was also sometimes added to *n* and *l* to indicate a palatal pronunciation (similar to the *ny* in English *canyon* and *li* in *scallion*), as today in Portuguese *vinho* 'wine' and *filho* 'son.' The palatal consonants *n* and *l* are also often depicted by doubled letters or other combinations of letters: the palatal *n* as *nn* (or its scribal variant *ñ*), *gn, nj,* or *in*; palatal *l* as *ll, gl, lj, il,* or *yl,* as well as the combinations *nh* and *lh,* already mentioned. Another device frequently used to stretch the capacity of the Latin alphabet was to distinguish the letters *i* and *j, u* and *v,* which were originally each single letters *i* (with variant form *j*) and *v* (with variant form *u,* and in Latin pronounced *u* or *w*). In Romance, *v* and *j* came to represent consonants, while *u* and *i* retained their vowel values.

Adding the letter *h*

The Latin letter *x,* an abbreviation for *ks,* was also put to other uses in Romance; in Portuguese, Catalan, Sicilian, and Old Spanish it represents a /sh/ sound, in modern Spanish a strong /h/ sound, more commonly spelled with a *j,* and in northern Italian dialects, the /z/ sound. Other letters pressed into use for new consonantal sounds were *z* (used in Italy for /ts/ and /dz/ sounds), Germanic *k* and *w,* and the Visigothic *ç* (for /ts/ and sometimes /s/, as today in French and Portuguese).

Vowels were less of a problem for early Romance scribes—diphthongs were simply shown as vowel combinations such as *ie, uo.* Later, the diaeresis (¨) was sometimes introduced to distinguish diphthongs from adjoining vowels that were to be pronounced separately. Non-Latin vowels are rarely clearly distinguished: French *u* (pronounced like German *ü*), for instance, was written *u* and not consistently distinguished from Latin *u* (pronounced as in English *lunar* and, in modern French, written *ou*). Nasal vowels in French are marked by a following *n* or *m*; in Portuguese a tilde (~) is often used for final nasal vowels and diphthongs (*ã, ãi*). Use of diacritics was not consistent until modern times; thus, so-called long and short *e,* still not always distinguished in Italian, are shown as *é* and *è* or *ê* (*e.g., élève* 'student') in French (since the 18th century) and as *e* and *é* in Portuguese (since about 1930). Romanian established the use of *î, â,* and *ă* only in the 20th century.

In most of the languages with a long history of writing, the original attempts by scribes at phonological transcription were followed by an "etymological" period in which Latinized spelling gained ground. Castilian was least subject to this fashion, and, because its phonology has changed comparatively little since the Middle Ages (when its spellings became more or less fixed), it has few orthographic problems today. Standard Italian retains a fairly etymological orthography that covers up various minor regional differences of pronunciation; small reforms have been made through the centuries (in the 17th century, for instance, the use of *h*—except in *ho, ha, hanno*—was discontinued; in the 20th century, *î* and *j* for *ii,* as in *studii,* have virtually disappeared), but chaos still reigns in the use of accents. Romanian suffered from etymologizing orthography in the 19th century, but successive edicts of the Romanian Academy, of which the most important was dated 1932, have established a more or less phonetic spelling (the notable exception being the depiction of final "soft" consonants by a following *i*). Modern Catalan, like other "minor" languages, has had the aid of expert linguists in the establishment of its orthography. A standard was proposed by Antonio María Alcover Sureda, a Catalan priest, philologist, and writer from Majorca, in 1913, which is accepted, with small variations, by most writers.

Only two of the Romance languages, French and Portuguese, have had major orthographic problems, mainly resulting from the radical transformations that have affected their phonology since the Latin period. Portuguese has attempted to overcome its difficulties by a series of governmental reforms during the 20th century, but, in spite of official agreements between Portugal and Brazil in 1931 and 1945, there is still little consistency in usage, with Brazilian writers, especially, remaining more conservative (*i.e.,* etymological). In France, in spite of vociferous demands for reform since the 16th century, only minor changes have been accepted (usually originally from unofficial sources, such as printers), so that French orthography today reflects 12th-century phonology, overlaid by the etymologizing of Middle French legal scribes. Battles still rage between the reformers, who deplore the absurdly large proportion of school time devoted to teaching spelling, and the defenders of tradition, who point out that the phonological character of French, with no consistent phonetic markers for the word, make it unsuitable for phonetic transcription and that written French has its own structure, not identical with that of the spoken language. (Re.P./M.S.)

Major orthographic problems

Germanic languages

The Germanic languages, a branch of the Indo-European language family, include a number of extinct languages as well as the earlier and present forms of German, Netherlandic, Afrikaans, English, Frisian, the Scandinavian languages, Yiddish, and their many dialects.

In numbers of native speakers, English, with 450 million, clearly ranks 4th among the languages of the world (after Mandarin, Hindi, and Spanish); German, with some 98 million, probably ranks 10th (after Bengali, Arabic, Portuguese, Russian, and Japanese). To these figures may be added those for persons with another native language who have learned one of the Germanic languages for commercial, scientific, literary, or other purposes. English is unquestionably the world's most widely used second language.

Table 14 presents information on each of the modern standard Germanic languages.

Table 14: Modern Standard Germanic Languages

	where spoken	approximate number of native speakers	use as a second language
English	Great Britain, Ireland, United States, Canada, Australia, New Zealand, South Africa	450,000,000	extreme
German	Germany, Austria, Switzerland (part)	98,000,000	extensive
Netherlandic (Dutch, Flemish)	The Netherlands, Belgium (part)	21,000,000	moderate
Swedish	Sweden, Finland (part)	8,000,000	slight
Afrikaans	South Africa (part), Namibia (part)	6,000,000	slight
Danish	Denmark	5,000,000	slight
Norwegian	Norway	4,000,000	slight
Frisian	The Netherlands, Germany	400,000	—
Yiddish	various countries	400,000	slight
Icelandic	Iceland	260,000	—
Faroese	Faroe Islands	44,000	—

The source for the English-, Netherlandic-, and Yiddish-language figures is B.E. Grimes (ed.), *Ethnologue* (1992); other figures, except Frisian, are from various official country sources.

The earliest historical evidence for Germanic is provided by isolated words and names recorded by Latin authors beginning in the 1st century BC. From approximately AD 200 there are inscriptions carved in the 24-letter runic alphabet. The earliest extensive Germanic text is the (incomplete) Gothic Bible, translated about AD 350 by the Visigothic bishop Ulfilas (Wulfila) and written in a 27-letter alphabet of the translator's own design. Later versions of the runic alphabet were used sparingly in England and Germany but more widely in Scandinavia—in the latter area down to early modern times. All extensive later Germanic texts, however, use adaptations of the Latin alphabet.

The names and approximate dates of the earliest recorded Germanic languages are listed in Table 15.

Table 15: Earliest Recorded Germanic Languages

	approximate dates, AD
Early Runic	200–600
Gothic	350
Old English (Anglo-Saxon)	700–1050
Old High German	750–1050
Old Saxon (Old Low German)	850–1050
Old Norwegian	1150–1450
Old Icelandic	1150–1500*
Middle Netherlandic	1170–1500*
Old Danish	1250–1500*
Old Swedish	1250–1500*
Old Frisian	1300–1500*

*Standard cutoff date for beginnings of modern Germanic languages.

The Germanic languages are related in the sense that they can be shown to be different historical developments of a single earlier parent language. Although for some language families there are written records of the parent language (*e.g.*, for the Romance languages, which are variant developments of Latin), in the case of Germanic no written records of the parent language exist. Much of its structure, however, can be deduced by the comparative method of reconstruction (a reconstructed language is called a protolanguage; reconstructed forms are marked with an asterisk). For example, a comparison of Runic *-gastiz,* Gothic *gasts,* Old Norse *gestr,* Old English *giest,* Old Frisian *iest,* and Old Saxon and Old High German *gast* 'guest' leads to the reconstruction of Proto-Germanic **gastiz.* Similarly, a comparison of Runic *horna,* Gothic *haurn,* and Old Norse, Old English, Old Frisian, Old Saxon, and Old High German *horn* 'horn' leads scholars to reconstruct the Proto-Germanic form **hornan.*

Reconstruction of parent language

Such reconstructions are, in part, merely formulas of relationship. Thus, the Proto-Germanic **o* of **hornan* in this position yielded *au* in Gothic and *o* in the other languages. In other positions (*e.g.*, when followed by a nasal sound plus a consonant) **o* yielded *u* in all the languages: Proto-Germanic **dumbaz,* Gothic *dumbs,* Old Norse *dumbr,* Old English, Old Frisian, and Old Saxon *dumb,* Old High German *tumb* 'dumb.' What may be deduced is that this vowel sounded more like *u* in some environments, but like *o* in others; it may be written as **u~o,* with the tilde indicating that it varied between these two pronunciations.

The above example shows that such reconstructions are more than mere formulas of relationship; they also give some indication of how Proto-Germanic actually sounded. Occasionally scholars are fortunate enough to find external confirmation of these deductions. For example, on the basis of Old English *cyning,* Old Saxon and Old High German *kuning* 'king,' the Proto-Germanic **kuningaz* can be reconstructed; this would seem to be confirmed by Finnish *kuningas* 'king,' which must have been borrowed from Germanic at a very early date.

By pushing the comparative method still farther back, it can be shown that Germanic is related to a number of other languages, notably Celtic, Italic, Greek, Baltic, Slavic, Iranian, and Indo-Aryan (Indic). All these language groups are subsequent developments of a still earlier parent language for which there are, again, no written records but which can be reconstructed as Proto-Indo-European (see above *Indo-European languages*).

LINGUISTIC CHARACTERISTICS OF THE PROTOLANGUAGE

The special characteristics of the Germanic languages that distinguish them from other Indo-European languages result from numerous phonological and grammatical changes.

Phonology. *Consonants.* Proto-Indo-European had 12 stop consonants: **p, *t, *k, *k^w; *b, *d, *g, *g^w; *bh, *dh, *gh,* and **g^{w}h;* and one sibilant, **s.* (Stops are produced with momentary complete stoppage of the breath stream at some point in the vocal tract.) By a change known as the Germanic consonant shift (or Grimm's law), the 12 stops changed in Germanic to voiceless fricatives, voiceless stops, and voiced fricatives, as illustrated in Table 16. A few examples: (1) Proto-Indo-European **p, *t, *k,* and **k^w,* as in Latin *piscis, tenuis, centum,* and *quod,* became Proto-Germanic **f, *þ, *x,* and **x^w,* as in English *fish, thin, hund(red),* and *what.* Proto-Germanic **x* and **x^w* early became **h* and **h^w* in some positions, giving the alternations of **h~x* and **h^w~x^w.* (2) Proto-Indo-European **d* and **g,* as in Latin *decem* and *genus,* became Proto-Germanic **t* and **k,* as in English *ten* and *kin.* (3) Proto-Indo-European **bh, *dh,* and **gh,* as in Sanskrit *bhū-, dhā-,* and *(g)hā-,* became Proto-Germanic **ƀ, *ð,* and **ǥ,* which later changed to the stops *b, d,* and *g* in some positions (*e.g.*, English *be, do,* and *go*), yielding **b~ƀ, *d~ð,* and **g~ǥ.* Proto-Indo-European **s,* as in Latin *sedeō,* was unchanged; Proto-Germanic kept **s,* as in English *sit.*

Grimm's law

Table 16: Sound Changes in the Germanic Consonant Shift

Proto-Indo-European voiceless stops	p	t	k	k^w
Proto-Germanic voiceless fricatives	f	þ	x	x^w
Proto-Indo-European voiced stops	b	d	g	g^w
Proto-Germanic voiceless stops	p	t	k	k^w
Proto-Indo-European voiced aspirated stops	bh	dh	gh	g^{w}h
Proto-Germanic voiced fricatives	ƀ	ð	ǥ	ǥ^w

In addition to these general changes, there were two special ones. (1) Proto-Indo-European **p, *t,* and **k* remained voiceless stops when preceded by **s* or another stop; thus, Proto-Indo-European **sp, *st, *sk, *pt,* and **kt* yielded Proto-Germanic **sp, *st, *sk, *ft,* and **xt,* respectively. For example, Proto-Indo-European **sp* and **st,* as in Latin *spuō* and *hostis,* remained **sp* and **st* in Proto-Germanic, as in English *spew* and *guest*; Proto-Indo-European **pt* and **kt,* as in Latin *captus* and *octō,* became Proto-Germanic **ft* and **xt,* respectively, in Old English *hæft* 'captured' and *eahta* 'eight.' (By still another change, Proto-Indo-European **tt* yielded Proto-Germanic **ss—e.g.,* Sanskrit *sattá-,* Old English *sess* 'seat.') (2) By a change known as Verner's law early Germanic voiceless **f, *þ, *x, *x^w,* and **s* (from Proto-Indo-European **p, *t, *k, *k^w,* and **s*) were voiced to **ƀ, *ð, *ǥ, *ǥ^w,* and **z,* respectively, when they followed an unaccented syllable, and the first four of these thereby merged with the already existing **ƀ, *ð, *ǥ,* and **ǥ^w* (from Proto-Indo-European **bh, *dh, *gh,* and **g^{w}h*). Thus, Proto-Indo-European **bhrā́tēr* became Proto-Germanic **brōþēr* (with *þ* after an accented syllable) and Old English *brōþor* 'brother'; but by Verner's law Proto-Indo-European **mātḗr* became Proto-Germanic **mōðēr* (with *ð* after an unaccented syllable) and Old English *mōdor* 'mother.' (Proto-Germanic **ð* yielded Old English *d*; the *th* of modern English *mother* is the result of a subsequent change.)

Verner's law

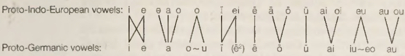
These changes yielded the following Proto-Germanic system of consonants: voiceless stops and fricatives, **p, *f, *t, *þ, *k, *h~x, *k^w, *h^w~x^w;* voiced stops and fricatives, **b~ƀ, *d~ð, *g~ǥ,* (**g^w~ǥ^w*); sibilants, **s, *z;* nasals, **m, *n;* liquids, **l, *r;* and semivowels, **w, *j(y).* The sound alternation of **g^w~ǥ^w* is parenthesized because it early became either **g~ǥ* or **w.* The sounds **k^w* and **h^w~x^w* occurred as such more or less clearly only in Gothic; elsewhere they became the sequences **kw* and **hw~xw,* or the labial element ^{w} was lost. All remaining consonants except *z* occurred between vowels both singly and doubly (*e.g.*, *-p-* and *-pp-*, *-t-* and *-tt-*).

Vowels. In addition to the above consonants (12 stops and the sibilant *s*), Proto-Indo-European also had vowels and resonants. The vowel of any given root was not necessarily fixed but varied in an alternation called ablaut. Thus, the root that means 'sit' was alternately **sed-, *sod-, *sd-, *sēd-,* and **sōd-* (English *sit* is from **sed-, sat* from **sod-,* and *seat* from **sēd-*); and the root that means *do* was **dhē-, *dhō-,* and **dhə-* (English *deed* is from **dhē-,* and *do* is from **dhō-*). Other Proto-Indo-European vowels were **a, *ā, *ī,* and **ū.* The Proto-Indo-European resonants, which functioned as vowels in some positions and as consonants in others, were **i, *u, *m, *n, *l,* and **r.* Thus, **bhr̥tó-* (Sanskrit *bhr̥tá-* 'borne') had syllabic **r̥* (*i.e.,*

Ablaut

r̥ functioning as a vowel), but **bhēreti* (Sanskrit *bhárati* 'he bears') had nonsyllabic **r* (*i.e., r* functioning as a consonant).

This Proto-Indo-European system of vowels contrasting with resonants was reshaped in Germanic by a number of changes. Syllabic **i*, **u*, **m̥*, **n̥*, **l̥*, and **r̥* became in Proto-Germanic the vowels **i* and **u* and the sequences **um*, **un*, **ul*, and **ur*, respectively; nonsyllabic **m*, **n*, **l*, and **r* developed into the nasals and liquids **m*, **n*, **l*, and **r*, respectively; nonsyllabic **i* and **u* before vowels resulted in the semivowels **j* (also symbolized as **y*) and **w*, though after vowels they continued to form diphthongs (**ei*, **ai*, **oi*; **eu*, **au*, **ou*). The Proto-Indo-European vowels and diphthongs then changed into Proto-Germanic sounds as follows:



In this diagram the lines between two sounds indicate that the Proto-Indo-European sound developed into the corresponding Proto-Germanic sound; for example, Proto-Indo-European **i* became either **i* or **e*, and Proto-Indo-European **ə*, **a*, and **o* coalesced in Proto-Germanic as **a*. These changes gave the following vowels for Proto-Germanic: short vowels, **i*, **e*, **a*, **u~o*; long vowels, **ī*, **ē²*, **ē¹*, **ū*, **ō*; diphthongs, **ai*, **au*, **iu~eo*. The origins of the vowel **ē²* are disputed, but it probably first arose from the diphthong *ei* in certain environments; **ē²* and **ē¹* were high and low midvowels, respectively. In Gothic the two *ē*'s merged, while elsewhere they remained distinct; thus, with **ē²*, Old High German *hiar* and Old Saxon, Old Norse, and Old English *hēr* 'here' but with **ē¹*, Old High German *tāt*, Old Saxon *dād*, Old Norse *dāð*, and Old English *dǣd* 'deed.' Proto-Germanic also had three nasalized vowels: long **ī̃*, **ā̃*, and **ū̃*, which arose when, in the sequences **inx*, **anx*, and **unx*, the *n* was lost with nasalization and lengthening of the preceding vowel.

Accent. Proto-Indo-European had a variable pitch accent that could fall on any syllable of a word, but in late Proto-Germanic, two changes occurred: first, the quality of the accent changed, such that articulatory energy was increasingly focused on the accented syllable; second, the position of the accent was regularized on the initial (root) syllable. These changes had far-reaching effects on the subsequent development of Germanic, for nonaccented syllables became subject to reduction and even total loss; thus, Proto-Germanic **kuningaz* but German *König*, Danish *konge*, English *king*. Reduction of unstressed vowels was often associated with the mutation or "umlaut" of preceding accented vowels. In some instances grammatical information that had been carried by suffixes came instead to be marked by alternations of root vowels—*e.g.*, **fōt*/**fōti* but English *foot/feet*, German *Fuss/Füsse*.

Grammar. *Declensions.* Proto-Germanic kept the Proto-Indo-European system of three genders (masculine, feminine, neuter) and three numbers (singular, dual, plural), though the dual was becoming obsolete. It reduced the Proto-Indo-European system of eight cases to six: nominative, accusative, dative, genitive, instrumental, and vocative, though the last two were becoming obsolete. In the adjective declensions there were two innovations: (1) To the Proto-Indo-European vowel types (**o-*, **ā-*, **i-*, and **u-* stems) it added some pronominal endings to give the Germanic "strong" adjective declension. (2) It extended the Proto-Indo-European **n-* stem endings to all adjectives to give the Germanic "weak" adjective declension. Contrast, in modern German, strong *gutes Bier* 'good beer' with weak *das gute Bier* 'the good beer.'

Conjugations. The Proto-Indo-European verb seems to have had five moods (indicative, imperative, subjunctive, injunctive, and optative), two voices (active and mediopassive), three persons (first, second, and third), three numbers (singular, dual, and plural), and several verbal nouns (infinitives) and adjectives (participles). In Germanic these were reduced to indicative, imperative, and subjunctive moods; a full active voice plus passive found only in Gothic; three persons; full singular and plural forms and dual forms found only in Gothic; and one infinitive (present) and two participles (present and past). The Proto-Indo-European tense-aspect system (present, imperfect, aorist, perfect) was reshaped to a single tense contrast between present and past. The past showed two innovations: (1) In the "strong" verb, Germanic transformed Proto-Indo-European ablaut into a specific tense marker (*e.g.*, Proto-Indo-European **bher-*, **bhor-*, **bhēr-*, **bhr̥-* in Old English *beran* 'bear,' past singular *bær*, past plural *bǣron*, past participle *boren*). (2) In the "weak" verb, Germanic developed a new type of past and past participle (*e.g.*, Old English *fyllan* 'fill,' past *fylde*, participle *gefylled*). Weak verbs fell into three classes depending on the syllable following the root (*e.g.*, Old High German *full-e-n* [from **full-ja-n*] 'fill,' *mahh-ō-n* 'make,' *sag-ē-n* 'say'). Gothic also had a fourth class: *full-nō-da* 'it became full.' Strong and weak verbs

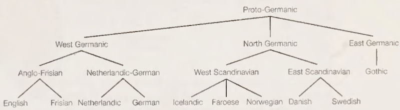
Many Proto-Germanic strong verbs showed a consonant alternation between **f* and **ƀ*, **þ* and **ð*, **x* and **g*, and **s* and **z* that was the result, through Verner's law, of the alternating position of the Proto-Indo-European accent. The forms in Table 17 illustrate changes resulting from Verner's law. In this particular word, English has generalized the **s* (now *z*): 'freeze,' 'froze,' 'frozen.' German has generalized the **z* (now *r*): *frieren, fror, gefroren.* And Netherlandic still shows the alternation: *vriezen, vroor, gevroren.* English has kept the alternation in only one verb: singular *was*, plural *were*. Traces of it still survive, however, in a few now isolated forms: *seethe* (Proto-Germanic **þ*) and its old past participle *sodden* (Proto-Germanic **ð*); *lose* (Proto-Germanic **s*) and its old past participle *(for)lorn* (Proto-Germanic **z*).

Table 17: Illustration of Verner's Law

	Proto-Indo-European	Proto-Germanic	Old English
'(to) freeze'	*préusonom	*freosanan	frēosan
'(it) froze'	*próuse	*fraus	frēas
'(they) froze'	*prusn̥t	*fruzun	fruron
'frozen'	*prusénos	*frozenaz	froren

An asterisk indicates an unattested, reconstructed form.

The emergence of Germanic languages. Like every language spoken over a considerable geographic area, Proto-Germanic presumably consisted of a number of geographic varieties or dialects that over time developed in different ways into the different early and modern Germanic languages. Late-19th-century scholars used a family tree diagram to show this splitting into dialects and the relationships among the dialects:



Though there is much truth in such a diagram, it overemphasizes the notion of "splits" into separate "branches" and obscures the fact that the transition from one dialect to another may be gradual rather than abrupt.

Mid-20th-century scholars, using the findings of archaeology and the methods of geographic linguistics, attempted to correct the distortions of this family-tree model by noting also the linguistic features shared by two or more dialect areas. Archaeological evidence suggests that about 750 BC a relatively uniform Germanic people was located in southern Scandinavia and along the North Sea and Baltic coasts from what is now The Netherlands to the Vistula River. By roughly 250 BC they had spread south, and five general groups are distinguishable: North Germanic in southern Scandinavia, excluding Jutland; North Sea Germanic, along the North Sea and in Jutland; Rhine-Weser Germanic, along the middle Rhine and Weser; Elbe Germanic, along the middle Elbe; and East Germanic, between the middle Oder and the Vistula rivers. Dialect similarities

By roughly AD 250 the division was much the same, though the Elbe group had spread southward to the

Danube River, and the East Germanic group had moved southeast into the Carpathian Mountains and beyond. Then, toward the end of the 4th century, began the great Germanic tribal migrations. North Sea Germanic speakers spread south along the coast of the Low Countries and began their conquest of Britain; North Germanic speakers moved into Jutland; the Rhine-Weser group (Franks) expanded farther into Gallo-Roman territory west of the Rhine; the Elbe group (Alemanni, Bavarians, and Langobardi [Lombards]) spread farther south to the Alps and beyond; and several East Germanic groups left the Oder-Vistula area to begin their wanderings.

This five-way division of Germanic peoples is based on archaeological evidence, and, while it agrees to some extent with deductions that can be made from the early linguistic evidence, the correspondence between archaeological and linguistic groupings is not completely straightforward.

The first major linguistic division that developed in Germanic was between East Germanic and Northwest Germanic. It can be dated roughly to the 1st–3rd centuries AD. Northwest Germanic is attested in the early runic inscriptions (*c.* AD 200–500) from Scandinavia and northern Germany and encompassed all the Germanic territory from Scandinavia southward across much of Germany and the Low Countries. Spoken over such a relatively large area, Northwest Germanic had at least minor dialectal distinctions from the start, but several linguistic innovations spread throughout.

The fragmentation of Northwest Germanic can be dated roughly to the period of the 3rd–6th centuries with the development of three major dialect divisions: in Scandinavia, North Germanic; in Jutland and the northwest of Germany, North Sea Germanic; and in central Germany, South Germanic. A number of linguistic developments from this period are shared by North Sea Germanic and South Germanic (but not by North Germanic), and the term West Germanic is used in recognition of the strong affinities between these two groups. During this same period, however, North Sea Germanic and North Germanic also sometimes follow similar lines with regard to phonological developments in contrast with South Germanic. The three-way division is then based in part on the differing participation of North Sea Germanic in shared developments with either North or South Germanic and in part on developments that were peculiar to each of the three dialects. North Sea Germanic was at once a transitional dialect and a centre of innovation within the larger Northwest Germanic dialect continuum.

In a position analogous to that of North Sea Germanic within Northwest Germanic, there surely existed transitional dialects between North Sea and South Germanic within West Germanic; the Old Saxon language may well reflect such a dialect, given that it shows a mixture of features agreeing with North Sea Germanic, as in English and Frisian, and with South Germanic, as in Old High German. The dialectal position of Netherlandic is complicated, and it is discussed further below. Old Saxon

Some linguistic features associated with these dialect divisions are the following:

1. Northwest Germanic versus East Germanic: In Northwest Germanic, the reduplicated forms found in the past tense of certain strong verbs were eliminated and instead

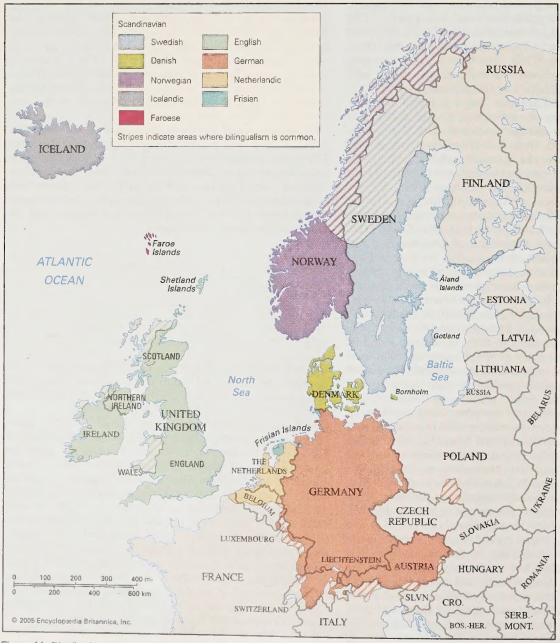


Figure 11: Distribution of the Germanic languages in Europe.

new ablaut alternations were employed (often involving the vowel $\bar{e}^2$), while, in East Germanic (Gothic), reduplicated forms were maintained—*e.g.,* Old Norse, Old English, Old Saxon *hēt,* Old High German *hiez* versus Gothic *haihait* 'was called.'

2. In East Germanic, Proto-Germanic **z* was maintained as *z* (or *s* in final position), but, in Northwest Germanic, **z* generally became *r*—*e.g.,* Proto-Germanic **maiz-* 'more,' Gothic *maiza* but Old Norse *meire,* Old English *māra,* Old High German *mēro.* Within Northwest Germanic, however, treatment of **z* in final position varied dialectally: (*a*) in North Germanic, it was maintained and appears as *r*; (*b*) in West Germanic (*i.e.,* North Sea and South Germanic), **z* was generally lost in final position in polysyllabic words; (*c*) in North Sea Germanic, **z* was also lost in final position in monosyllabic words; for example, Proto-Germanic **dagaz* 'day,' Gothic *dags,* Old Norse *dagr,* Old High German *tag,* Old Saxon *dag,* Old English *dæg;* Proto-Germanic **wīz* or **wiz* 'we,' Gothic *weis,* Old Norse *vēr,* Old High German *wir,* Old Saxon *wī,* Old English *wē.*

3. Associated with the West Germanic grouping are: (*a*) the change of **ð* to *d* in all positions—*e.g.,* Proto-Germanic **blōð-* 'blood,' Old Norse *blōð-* but Old English and Old Saxon *blōd,* Old High German *bluot*; (*b*) the replacement of the inherited second person singular past tense forms of strong verbs—*e.g.,* Proto-Germanic **namt* 'you took,' Old Norse *namt* but Old English *nōme,* Old Saxon and Old High German *nāmi.*

4. Some changes were shared across major dialects but carried out to markedly differing degrees or at different times. For example, in West Germanic all consonants (except *r*) preceded by a short vowel and followed by *j* were geminated (doubled), but in North Germanic only velars (*k, g*) were affected: compare Proto-Germanic **satjanan* 'set,' Gothic *satjan,* Old Norse *setja* but Old English *settan,* Old Saxon *settian,* Old High German *setzen*; Proto-Germanic **lagjanan* 'lay,' Gothic *lagjan* but Old Norse *leggja,* Old English *lecgan,* Old Saxon *leggian,* Old High German *lecken.* Similarly, the loss of nasals in the cluster *nx* with lengthening and nasalization of a preceding short vowel occurred in Proto-Germanic; in North Germanic this change also occurred in sequences of a short vowel plus *ns* (compare Old Norse *oss* 'us,' Old High German *uns*) and in North Sea Germanic when the nasal was followed by any voiceless fricative (*mf, nþ, ns*; compare English below). In addition, fronting of nonfront vowels under the influence of following *i* or *j* in unaccented syllables, "*i*-umlaut," developed earlier (6th–7th centuries) and more consistently in North Sea Germanic and North Germanic than in South Germanic (8th–9th centuries).

EAST GERMANIC

The East Germanic languages are all long extinct; they descended from the dialect group that developed between the middle Oder and the Vistula rivers.

History. According to historical tradition, at least some of the Germanic tribes migrated to the mouth of the Vistula from Scandinavia. Little is known of Gepidic, Rugian, and Burgundian; some knowledge of Vandalic, Visigothic, and, especially, Ostrogothic is provided by the names recorded in Greek and Latin writings. The only East Germanic language on which there is extensive information is the Gothic—more specifically, Visigothic—that was spoken along the western shore of the Black Sea about the middle of the 4th century AD.

Gothic language

Knowledge of Gothic is derived primarily from the remains of a Bible translation made for the Visigoths living along the lower Danube by Ulfilas, a Visigothic bishop of the Arian church, who lived during the 4th century. The surviving manuscripts of this translation, which are not originals but later copies thought to have been written in northern Italy during the period of Ostrogothic rule (493–554), include considerable portions of the New Testament. The best-known manuscript is the Codex Argenteus, written in silver and gold letters on purple parchment and containing (in 188 leaves remaining from an original 330 or 336) portions of the four Gospels. Closely related to these biblical manuscripts are eight leaves containing fragments of a commentary (called the *Skeireins* in Gothic) on the Gospel According to John. Minor nonbiblical texts include a fragment of a calendar, two deeds containing some Gothic sentences, and a 10th-century Salzburg manuscript that gives the Gothic alphabet, a few Gothic words with Latin transliteration, and some phonetic remarks with illustrative examples.

In the 4th and 5th centuries Gothic (Visigothic and Ostrogothic) must have spread to some degree, along with the conquering Goths, throughout much of southern Europe; but there is no evidence for its survival in Italy after the fall of the Ostrogothic kingdom, and in Spain it is doubtful that the Visigoths retained their language until the Arab conquest. In the 9th century the German monk Walafrid Strabo mentions that Gothic was still being used in some churches near the lower Danube. After that time Gothic seems to have survived only among the Goths of the Crimea, who were last mentioned in the middle of the 16th century by a Flemish diplomat named Augier Ghislain de Busbecq, who, while on a mission to Constantinople in 1560–62, collected a number of words and phrases showing that their language was still essentially a form of Gothic.

Gothic alphabet

Characteristics. The Gothic alphabet, said to have been created by Ulfilas, contained 27 symbols, two of which functioned only as numbers, while the remaining 25 were used as both numbers and letters. The shape, numerical value, and ordering of the symbols show clearly that the alphabet was based primarily on that of Greek, though a few symbols seem to have been adapted from the Latin alphabet.

Phonology. The Gothic consonant system seems to have been largely identical with that assumed above for Proto-Germanic: *p, t, k,* k^w (this last sound was probably much like the *qu* in *queen*); *f, þ, h,* h^w (this last sound was probably pronounced much like the *wh* in *white*); *b, d, g; s, z; m, n; l, r; w, j.* The nasal *n* was presumably velar before the velar consonants *k, q,* and *g;* in these positions it was usually written (as in Greek) as *g* or *gg.* Examples of this spelling include *dragk* 'drank,' *igqis* 'you two,' and *briggan* 'bring,' although *n* was occasionally used as in Latin (*e.g., þank* 'thanks,' *inqis* 'you two,' and *bringiþ* 'bring ye').

The Gothic alphabet contained the five simple vowel symbols, *i, e, a, o,* and *u,* from which four compound symbols, *ei, ai, au,* and *iu,* also were made; in addition, *w* was used to transliterate Greek υ and οι (both of which were pronounced as umlauted *u* /ü/ in 4th-century Greek). The generally accepted development of the Proto-Germanic vowels in Gothic can be diagrammed as follows:

Proto-Germanic:	i e	a	o ~ u	i, ī	ē², ē¹	ai	ā̃	au	ō	ū, ũ̄	eo ~ iu
Gothic:	i ɛ	a	ɔ u	ī	ē	ɛ̄	ā	ɔ̄	ō	ū	iu
Spelling:	i ai	a	au u	ei	e	ai	a	au	o	u	iu

Brackets in the Proto-Germanic line indicate that the two linked sounds coalesced into one; brackets in the Gothic line indicate two variants of the same sound that are in different phonetic environments. Proto-Germanic **i* and **e* apparently first merged as a single vowel and then became Gothic *i* in most positions but became *ai* before *h,* h^w, and *r.* Similarly, Proto-Germanic **u~o* became Gothic *u* in most positions, but *au* before *h,* h^w, and *r.*

Archaic features. Gothic shows a number of archaic features that had been almost or entirely lost by the time the other Germanic languages began to appear in writing; among these are a passive voice and one type of past tense formed with reduplication, a dual number in the first and second persons of its verbs and pronouns, and a special vocative case in one noun class. At the same time, Gothic also shows changes from Proto-Germanic, among which are the shortening of most long vowels in final unstressed syllables and the loss of most short vowels (*e.g.,* Proto-Germanic **erþō* 'earth' became Gothic *airþa,* Proto-Germanic **stainaz* 'stone' became Gothic *stains*). Finally, voiced fricatives that occurred or came to occur at the end of a word are unvoiced (*e.g.,* nominative **hlaiƀaz,* accusative **hlaiƀan* 'bread, loaf' changed to *hlaifs* and *hlaif,* respectively [but dative *hlaiba*]).

WEST GERMANIC

The West Germanic languages are those that developed from the North Sea and South Germanic groups mentioned above. Out of the many local West Germanic dialects the following six modern standard languages have arisen: English, Frisian, Netherlandic (also called Dutch, Flemish), Afrikaans, German, and Yiddish.

English and Frisian changes

English. English and Frisian are descended from North Sea Germanic. The most striking changes that distinguish them from the other Germanic languages are the loss of nasal sounds before the Proto-Germanic voiceless fricatives _*f, *þ,_ and _*s_ (contrast the following pairs of words, in which English loses the nasal but German preserves it: before _f—soft/sanft;_ before _þ—other/ander;_ before _s—us/uns, goose/Gans_); palatalization of Proto-Germanic _*k_ before front vowels and _*j,_ giving modern English _ch_ (English/German pairs: _chin/Kinn, birch_ [Old English _birce_]/_Birke_); and palatalization of Proto-Germanic _*g_ before front vowels, giving modern English _y_ (English/German pairs include _yield/gelten, yester-_[_day_]/_gestern, yard_ [Old English _geard_]/_Garten;_ this palatalized _g_ merged with the _j_ [_y_ sound] from Proto-Germanic _*j: year/Jahr_).

Other changes include palatalization of _gg_ before _j_ to Old English _cg_ (Proto-Germanic _*brugjō,_ pre-Old English _*bruggju,_ Old English _brycg_ 'bridge'; contrast the unpalatalized _ck_ from _gg_ of German _Brücke_ 'bridge'); a front reflex of Proto-Germanic _*ē_^{1} (English/German pairs include _deed/Tat, seed/Saat, sleep/schlafen, meal/Mahl_); and backing and raising of nasalized _ā,_ from Proto-Germanic _*ā_ and _*a_ before nasal plus _f, þ,_ and _s_ (English/German pairs include _brought/brachte, thought/dachte, other/ander,_ and _goose/Gans_).

For further information on English, see below _English language_.

Frisian. A thousand years or so ago Frisian was apparently spoken throughout a North Sea coastal area extending from the modern Netherlands province of Noord-Holland (North Holland) on up to modern German Schleswig and the adjacent offshore islands. During the following centuries the Frisian of much of this area was gradually replaced by local Netherlandic and Low German dialects, so that Modern Frisian is now spoken in only three remaining areas: (1) West Frisian, in the Netherlandic province of Friesland, including the island of Schiermonnikoog and two-thirds of the island of Terschelling (altogether some 400,000 speakers), (2) East Frisian, in the German Saterland (some 1,000 speakers; this area was apparently settled in the 12th or 13th century from the former East Frisian area to the north), and (3) North Frisian, along the west coast of German Schleswig and on the offshore islands of Sylt, Föhr, Amrum, the Halligen, and Helgoland (altogether some 8,000 speakers).

History. The earliest manuscripts written in Frisian date from the end of the 13th century, though the legal documents that they contain were probably first composed, in part, as early as the 11th century. This stage of the language, until about 1575, is known as Old Frisian. The last written document of this period dates from 1573, after which Frisian was relatively little used as a written language for some three centuries.

From the start Old Frisian shows all the features that distinguish English and Frisian from the other Germanic languages. These include loss of the nasal sound before Proto-Germanic _*f, *þ,_ and _*s_ (_e.g.,_ Proto-Germanic _*fimf, *munþ-,_ and _*uns_ became Old Frisian _fīf_ 'five,' _mūth_ 'mouth,' and _ūs_ 'us'), palatalization of Proto-Germanic _*k_ before front vowels and _*j_ (_e.g.,_ Proto-Germanic _*kinn-_ and _*lē_^{1}_kj-_ became Old Frisian _tzin_ 'chin' and _lētza_ 'physician' [compare English archaic _leech_]), and palatalization of Proto-Germanic _*g_ before front vowels (_e.g.,_ Proto-Germanic _*geldan-_ became Old Frisian _ielda_ 'yield'). This merged with the _j_ from Proto-Germanic _*j,_ as in Proto-Germanic _*jē_^{1}_r-_ or Old Frisian _iēr_ 'year.' In addition, Old Frisian shows palatalization of _gg_ from Proto-Germanic _*g_ before _j_ (_e.g.,_ Proto-Germanic _*lagjan-,_ with doubling _*laggjan,_ became Old Frisian _ledza_ 'to lay'); a front vowel for Proto-Germanic _*ē_^{1}, as in Proto-Germanic _*dēð,_ Old Frisian _dēd;_ and backing and raising of nasalized _ā_ from Proto-Germanic _*ā_ and Proto-Germanic _*a_ before nasal plus _*f, *þ, *s,_ as in Proto-Germanic _*brāxt-, *anþar-,_ and _*gans-,_ which became Old Frisian _brocht_ 'brought,' _ōther_ 'other,' and _gōs_ 'goose.'

About the beginning of the 19th century it appeared that the age-old replacement of Frisian by Netherlandic and Low German would continue unabated and that the language would soon become extinct. But with 19th-century Romanticism a new interest in local life arose, and societies were formed for the preservation of the Frisian language and culture. Very slowly, the aims of this "Frisian movement" came to be realized, especially in the Netherlands province of Friesland, where in 1937 Frisian was accepted as an optional course in elementary schools; a Frisian Academy was founded in 1938; and in 1943 the first Frisian translation of the Bible was published. In 1955 Frisian was approved as the language of instruction in the first two years of elementary school (though only about one-fourth of all schools use it in this way), and in 1956 the use of Frisian in courts of law was approved.

Status of Frisian

Despite this gradual reemergence of Frisian, Netherlandic still functions as the primary standard language of Friesland. Nearly all school instruction is given in Netherlandic; all daily newspapers are printed in Netherlandic (though they contain occasional articles in Frisian); and the majority of television and radio broadcasts are in Netherlandic. There is a small and enthusiastic Frisian literary movement, but its works are not widely read. Furthermore, though Frisian continues to be widely used as the language of everyday oral communication, it is increasingly a "Netherlandic" Frisian, with numerous borrowings from standard Netherlandic.

The status of Frisian in the East and North Frisian areas of Germany is far more tenuous. There German performs all the functions of a standard language, and Frisian serves only as yet another local dialect, comparable to the many surrounding local dialects of Low German. No standard North Frisian or East Frisian exists.

Characteristics. The following remarks refer to the more or less standard West Frisian that is developing in the province of Friesland.

Frisian has the following system of consonants, given here in the usual spellings: stops, _p, b, t, d, k, g;_ fricatives, _f, v, s, z, ch, g;_ nasals, _m, n, ng;_ liquids, _l, r;_ and glides, _w, h, j._ Examples (given here in part to show the close relationship between Frisian and English) include _p, t,_ and _k_ (unaspirated) in _peal_ 'pole,' _twa_ 'two,' and _kat_ 'cat'; _b, d,_ and the stop symbolized by the letter _g_ in _boi_ 'boy,' _dei_ 'day,' and _goed_ 'good'; _f, s,_ and _ch_ in _fiif_ 'five,' _seis_ 'six,' and _acht_ 'eight'; _v, z,_ and the fricative symbolized by the letter _g_ in _tolve_ 'twelve,' _tûzen_ 'thousand,' and _wegen_ 'ways'; _m, n,_ and _ng_ in _miel_ 'meal,' _need_ 'need,' and _ring_ 'ring'; _l_ and _r_ in _laem_ 'lamb' and _reep_ 'rope'; _w, h,_ and _j_ in _wy_ 'we,' _hy_ 'he,' and _jo_ 'you.' As the final letter of a word, voiced _b, d, z,_ and _g_ are generally unvoiced to _p, t, s,_ and _ch_.

Frisian has the following system of stressed vowels and diphthongs. The symbols given in Table 18 refer to the actual sounds rather than to Frisian spellings, which are often irregular: Frisian also has an unstressed vowel _ə_ (pronounced as the _a_ in English _sofa_), which occurs only in unstressed syllables.

Table 18: The Vowel System of Frisian

short vowels			long vowels			rising diphthongs			falling diphthongs		
i	ü	u	ī	ǖ	ū		ui		iə	yə	uə
I	ö	o	ē	ȫ	ō	ɛi	oəi	öü	Iə	öə	oə
ɛ		ɔ	ɛ̄		ɔ̄	ai	āi	au			
	a			ā							

Dialects of Friesland

Dialects. The Frisian dialects of the Netherlands province of Friesland are, with three exceptions, relatively uniform, though it is customary to make a distinction between Wouden Frisian in the east, Klei Frisian in the west (the variety on which standard Frisian is largely based), and Southwest Corner Frisian in the southwest. The three exceptions are the island dialect of East and West Terschelling and the dialects of the city of Hindeloopen and of the island of Schiermonnikoog. These latter two differ so greatly that they are not intelligible to other speakers of West Frisian and are both dying out. Quite different from any of these is the so-called City Frisian (Stedfrysk, or Stedsk) spoken in the cities of Leeuwarden, Franeker, Harlingen, Bolsward, Sneek, Staveren, and Dokkum. Despite the name, this is not Frisian at all but a variety of Netherlandic strongly influenced by Frisian. Similar in nature are the dialects of Heerenveen and Kollum, of the middle section of the island of Terschelling, and of Het Bildt (a coastal area northwest of Leeuwarden, diked in and settled by Hollanders during the 16th century).

East Frisian survives today only in the German Saterland, consisting of the three parishes of Ramsloh, Strücklingen, and Scharrel, each with a slightly different dialect. The area to the north is called East Frisia (German Ostfriesland), and the local dialect East Frisian (German Ostfriesisch), although it is actually not Frisian but the local variety of Low German.

Though North Frisian is spoken in only a small geographic area by only some 8,000 persons, it exists in an extraordinary number of local dialects, some of which are mutually unintelligible. Because of this, it would be almost impossible to develop a single standard North Frisian that could be used throughout this area. North Frisian dialects are customarily divided into Insular North Frisian (Sylt, Föhr-Amrum, Helgoland) and Continental North Frisian (the Halligen Islands and the coast of Schleswig), the latter in seven main varieties and further subvarieties. Because this linguistic area long bordered on Danish, it was extensively influenced by the neighbouring Danish dialects. In more recent times it has been heavily influenced by German, both standard German and the neighbouring Low German dialects. Today all speakers of North Frisian are bilingual or trilingual; all learn Frisian at home and standard German in school, and many also learn dialectal Low German.

Netherlandic (Dutch, Flemish). Netherlandic is the national language of The Netherlands and one of the two national languages (besides French) of Belgium. Popular English usage applies the term Dutch to the Netherlandic of Holland and the term Flemish to the Netherlandic of Belgium, but in fact they are one and the same standard language. In its various forms, standard and dialectal, Netherlandic is the indigenous language of most of The Netherlands (all but the Frisian-speaking province of Friesland), of northern Belgium, and of a small part of France immediately to the west of Belgium. It also is used as the language of administration in the Dutch dependency of the Netherlands Antilles and in Suriname, a former dependency. A derivative of Netherlandic, Afrikaans, is one of the national languages of South Africa.

Identity of "Dutch" and "Flemish"

As a written language, Netherlandic is quite uniform; it differs in The Netherlands and Belgium no more than written English does in the United States and Great Britain. As a spoken language, however, it exists in far more varieties than does the English of North America. At one extreme is Standard Netherlandic (Algemeen Beschaafd Nederlands, 'General Cultured Netherlandic'), which is used for public and official purposes and is the language of instruction in schools and universities. It is everywhere quite uniform, though speakers usually show by their accents the general area from which they come. At the other extreme are the local dialects, used among family and friends and with others from the same village.

History. Netherlandic emerged as a structurally distinct branch of West Germanic as the result of language contact between speakers of North Sea Germanic and speakers of the South Germanic "Franconian," or Frankish. The crucial early period of this contact occurred in the 7th and 8th centuries and resulted from the expansion of Frankish (Merovingian and early Carolingian) power into the western coastal areas that were populated by North Sea Germanic groups. The most important structural characteristic of Netherlandic is its strikingly anomalous development of *i*-umlaut; whereas in all other West and North Germanic languages, *i*-umlaut affected all nonfront vowels, in standard Netherlandic only *ă* shows umlaut fronting. In the dialects this limited development of umlaut is found only in the coastal areas (Flanders, Zeeland, Holland); eastern dialects show umlaut developments like those of the neighbouring German dialects and standard German.

The structural peculiarity of the coastal dialects can best be explained as the result of the imperfect acquisition of Frankish by North Sea Germanic speakers in the period about AD 700. The resulting new form of Frankish, which was structurally affected by contact with North Sea Germanic, first developed in centres such as Ghent and Bruges, both of which gradually rose to great economic and cultural prominence. In the course of the Middle Ages, the local North Sea Germanic dialects of the west were replaced by the restructured Frankish—*i.e.,* Netherlandic; at the same time, Netherlandic influence also spread eastward (see *Dialects* below).

Little vernacular material from the Low Countries has survived from the period before 1200. There are, for example, glosses of psalms dating to the 10th or 11th century in a dialect from the area around the modern Dutch-German border. Beyond this, only a few sentences and words are recorded in various texts, but a large body of onomastic material also exists. About 1200, the Middle Dutch period begins and from this time exists a large corpus of both literary and nonliterary documents. Though texts from all dialect areas have survived, the texts from Flanders are by far the most numerous.

The economic and cultural preeminence of the Flemish cities was especially great during the Middle Dutch period, and both scribal and linguistic influences from Flanders can be seen in documents originating in other regions. In the 15th century the cities of Brabant began to surpass the Flemish cities in importance. In turn, partly because of the failure of the southern provinces to break free of Spanish rule, preeminence shifted again in the late 16th century to Amsterdam and the other cities of Holland. It was in Holland during the 17th century that a genuine standard variety arose, yet, in the written language, strong Flemish and Brabant influences remain. Under foreign rule the language lost status in the southern provinces and in the 18th and 19th centuries was relegated to the position of a *lingua rustica,* with French serving as the standard language. This situation was rectified through political action during the 20th century.

Characteristics. Modern Standard Netherlandic has the following consonants, given here in the usual spellings: stops, *p, b, t, d, k;* fricatives, *f, v, s, z, ch, g;* nasals, *m, n, ng;* liquids, *l, r;* glides, *w, h, j.*

The voiced stops and fricatives *b, d, v, z,* and *g* are unvoiced to *p, t, f, s,* and *ch,* respectively, as final letters of a word. The spelling shows this in the case of *v* and *z* (plural *dieven* 'thieves,' *huizen* 'houses,' but singular *dief* 'thief,' *huis* 'house') but does not show it in the case of *b, d,* and *g* (plural *ribben* 'ribs,' *bedden* 'beds,' *dagen* 'days,' but singular *rib* 'rib,' *bed* 'bed,' *dag* 'day,' pronounced "rip," "bet," "dach").

Netherlandic has three classes of vowels and diphthongs: (1) six checked vowels, which are short and always followed by a consonant, (2) 10 free vowels and diphthongs, most of them usually long, which need not be followed by a consonant, and (3) a vowel that occurs only in unstressed syllables. They form the system shown in Table 19 (the traditional spelling is to the left, and to the right is a notation, used by some linguists, that indicates the distinctive sounds [phonemes] of the language). Unlike the English spelling system, which in its basic design has remained essentially unchanged since the days of Chaucer (d. 1400), the Netherlandic spelling system has undergone a series of official reforms to keep it in line with changes in pro-

Table 19: Vowel System of Netherlandic

traditional spelling						linguistic notation					
checked			free			checked			free		
			ie	uu	oe				ī̆	ṻ̆	ū̆
i	u	o	ee	eu	oo	ɪ	ö̆	o	ē	ȫ	ō
e		o	ij, ei	ui	ou, au	ɛ		ɔ	ɛi	ɔ̈ü	ɔu
	a			aa			a			ā	
unstressed: e						unstressed: ə					

Netherlandic spelling

nunciation. The principal inconsistencies in the spelling of vowels are the spellings *ij* and *ei,* which both symbolize the same diphthong, pronounced somewhat between the *ai* of English *aisle* and the *ai* of English *maid* (*bijt* 'he bites' rhymes with *feit* 'fact'), and the spellings *ou* and *au,* which both symbolize the same diphthong, pronounced somewhat between the *ow* of English *now* and the *ow* of English *low* (*bouw* 'building' rhymes with *nauw* 'narrow'). Free vowels are written with double letters in closed syllables (*vuur* 'fire,' *boot* 'boat'), but with single letters in open syllables (*vuren* 'fires,' *boten* 'boats'). In contrast the checked vowels are always written with single letters.

Dialects. There is no structurally significant linguistic boundary that coincides with the national and standard language boundaries between The Netherlands and Belgium on the one side and Germany on the other; at the level of local dialects, the entire Netherlandic-German territory from the English Channel to the Alps constitutes a continuum, with only gradual transitions from one village to the next.

The most important dialect boundaries in Netherlandic territory run from north to south and separate the area into a coastal group (Flanders, Zeeland, Holland), with the anomalous development of umlaut and sundry North Sea Germanic features; a central area (Brabant, Utrecht), where umlaut developed as in German but, under influence from the coastal dialects, was eliminated from morphological marking; and an eastern area (Limburg, eastern North Brabant, Gelderland), where umlaut alternations are still used for morphological marking. These dialects have traditionally been called "Frankish"; the dialects of the northeastern part of The Netherlands (Overijssel, Drenthe, Groningen) have been called "Saxon" and show certain affinities with Low German dialects to the east. On the basis of other linguistic features, it is also possible to group together the dialects to the south and to the north of the Rhine and Meuse rivers.

Pressure from the standard language has made traditional local dialects in The Netherlands extinct or moribund in many areas. In Belgium, however, the dialects have survived to a far greater degree.

Afrikaans. Afrikaans is one of the official languages of South Africa, where it is the native language of roughly equal numbers of whites and nonwhites (especially "Coloureds"). Few languages have engendered as much controversy, with regard to both historical development and place in modern society.

History. Afrikaans is derived from a colonial dialect of Netherlandic ("Cape Dutch"). The Dutch presence in southern Africa began in 1652, when the Dutch East India Company took the first steps in establishing a station at the Cape of Good Hope. The European colonial population grew slowly, and only about half was of Dutch origin; in addition, there were considerable numbers of Germans and French. The mixing of Netherlandic dialects with the languages of non-Dutch Europeans has often been cited as a central factor in the divergence of Afrikaans from Dutch, but this view is linguistically flawed; among a highly similar European colonial population in the New Netherland colony in North America, there also developed a colonial dialect, but one that deviated little from the dialects of Holland and Utrecht provinces.

The crucial factor in the formation of Afrikaans was the development of a creolized variety of Cape Dutch among both indigenous Khoisan peoples and the imported slave population of the colony. As a result of the intimate interaction of a part of the European population with the Khoisan and slaves, a variety of the colonial dialect arose. This variety contained many features that had first arisen in the creolized variety. Afrikaans is then a product of a cross between the colonial dialect itself and a creolized variety of that dialect and is therefore not considered a creole but rather a partially creolized language. To this day, however, varieties exist, especially among nonwhite groups, that display many more creolelike features than the standard does. Standard Afrikaans is lexically extremely close to Netherlandic but has a markedly simplified morphology (*e.g.,* lost are person and number markings on verbs, strong verb root alternations, and nominal gender) and a number of syntactic innovations (*e.g.,* double-negation with mandatory clause-final *nie*).

Both the colonial dialect and the creolized variety developed in the first half-century or so of the colony's history, but Afrikaans-like constructions are first attested—only sporadically—from the late 18th and early 19th centuries. About the middle of the 19th century, the effort to make Afrikaans a literary language began. It came gradually to be used in newspapers. It was adopted for use in schools in 1914 and was accepted for use in the Dutch Reformed Church in 1919. In 1925 the South African Parliament declared it to be an official language, replacing Netherlandic, and Afrikaans was retained as an official language in the new constitution of 1996.

Characteristics. Afrikaans has the following consonants, given here in the conventional spellings: stops, *p, b, t, d, k, gh/g;* fricatives, *f/v, w, s, z, g;* nasals, *m, n, ng;* liquids, *l, r;* glides, *h, j.* There are numerous differences between Afrikaans and Netherlandic. Netherlandic *-g-* (*-gg-*) is a voiced fricative, but Afrikaans *-g-* (*-gg-*) is in many instances a voiced stop. Unlike Netherlandic, Afrikaans also has this voiced stop initially in a few loanwords. Netherlandic has voiced fricatives initially (*v-, z-, g-*); corresponding words have voiceless initial fricatives in Afrikaans. Afrikaans, however, has voiced *z-* in loanwords: *Zoeloe* 'Zulu.' Netherlandic has initial *s* plus fricative *ch* as in *schoen* 'shoe'; corresponding words have *s* plus *k* in Afrikaans: *skoen.* Netherlandic has *-ft, -st,* and *-cht* as in *gift* 'poison,' *nest* 'nest,' *nacht* 'night'; corresponding words show loss of *-t* in Afrikaans: *gif, nes,* and *nag.*

Afrikaans has the system of vowels shown in Table 20 (usual spelling to the left; notation used by linguists to indicate distinctive sounds to the right). As in Netherlandic, *uu, ee, oo,* and *aa* are written with single letters in open syllables, and single consonant letters are doubled in open syllables to show that the preceding vowel is short.

German. German is spoken throughout a large area in central Europe, where it is the national language of Germany and of Austria and one of the four national languages (beside French, Italian, and Romansh) of Switzerland. From this homeland it has been carried by emigration to many other parts of the world; there are German-speaking communities in North and South America, South Africa, and Australia.

Dispersion of German

As a written language German is quite uniform, differing in Germany, Austria, and Switzerland no more than written English does in the United States and the British Commonwealth. As a spoken language, however, German exists in far more varieties than English. At one extreme is Standard German (Hochsprache), based on the written form of the language and used in radio, television, public lectures, the theatre, schools, and universities. It is relatively uniform, although speakers often reveal regional

Table 20: Vowel System of Afrikaans

usual spelling									linguistic notation								
short vowels			long vowels			diphthongs			short vowels			long vowels			diphthongs		
ie	uu	oe	ie	uu	oe	ei/y			i	y	u	ī	ȳ	ū	əi		
	i			î		ui	ou	oo/o		i			ī		œi	œu	oə*
	i/e		ê	û	ô		eu			ə		ɛ̄	œ̄	ɔ̄		øə*	
e	u	o		aa/a		ee/e		ai	ɛ	œ	ɔ		ā		eə*		ai
	a									a							

*The spellings *ee, eu,* and *oo* are generally pronounced as diphthongs (eə, øə, oə) though prescriptive grammars still indicate that they are long monophthongs (ē, œ̄, ō).

accents. At the other extreme are the local dialects, which differ from village to village. Between these two extremes there is a continuous scale of speech forms; in cities these forms are often close to the standard language and are called Colloquial German (Umgangssprache).

History. From the point of view of the modern local dialects, the territory within which German and Netherlandic are spoken is a single speech area. It is possible to travel from Austria, northern Italy, and much of Switzerland into Germany, eastern France (Alsace and part of Lorraine), Luxembourg, northern Belgium, and The Netherlands without encountering a village where the local speech is suddenly different. The only sharp breaks occur when one enters the French-speaking parts of France and Belgium or the Frisian-speaking parts of The Netherlands and Germany.

The most striking dialect differences within this large area are those that divide Netherlandic–Low German in the lowlands of the north from High German in the highlands of the south. When the Germanic tribes migrated into southern Germany during the early centuries AD, their speech had the voiceless stops *p, t,* and *k* in much the same distribution as in modern English. Then, probably during the 6th century, there occurred a change customarily called the High German consonant shift. At the beginning of words and when doubled, *p, t,* and *k* came to be pronounced as affricates; after a vowel they came to be pronounced as long fricatives. The modern results, compared with related English words, are shown in Table 21.

High German consonant shift

Table 21: Results of the High German Consonant Shift

p-	*p*ound	*Pf*und	*pp*	ap*p*le	A*pf*el	*Vp*†	ho*p*e	ho*ff*en
t-	*t*en	*z*ehn	*tt*	si*tt*ing	si*tz*en	*Vt*†	bi*t*e	bei*ss*en
k-	*c*an	*kh*ann*	*kk*	li*ck*	le*kch*en*	*Vk*†	ma*k*e	ma*ch*en

**Khann* and *lekchen,* with affricates, are southern dialect forms; standard German has stops: *kann, lecken.* †*V* represents any vowel.

These changes occurred in the south of the German speech area and then spread north, some extending farther than others. The situation at the end of the 19th century was as indicated in Figure 12. Line 2, *maken/machen,* is generally chosen as the boundary between Low German and High German, because it is typical for the shift of *p, t,* and *k* after vowels to *ff, ss,* and *ch,* respectively (*hopen/hoffen, bīten/beissen, maken/machen*), and of *t* and *tt* to *z* and *tz,* respectively (*ten/zehn, sitten/sitzen*). The shift of *ik* 'I' to *ich* is indicated by line 1, which shows that the shift of *k* to *ch* after a vowel in this particular word spread unusually far. Line 3, which indicates the shift of *Dorp* 'village' to *Dorf* (compare archaic English *thorp*), shows that shifted *p* after *r* and *l* did not spread as far north as the shifted *p, t,* and *k* after a vowel. And line 4, indicating the shift of *dat* 'that' to *das,* shows that the shift of *t* to *s* after a vowel spread still less far north in this word (and in a few others: *it/es* 'it,' *wat/was* 'what'). The striking way in which these lines fan out in the west (in the area along the Rhine River) has led to their being referred to as the Rhenish fan.

Figure 12: The Netherlandic-German dialect divisions. Numbers refer to isoglosses described in text.

The shift of *p* when doubled or at the beginning of a word occurred in a much smaller area. Line 5, showing the shift of *Appel* 'apple' to *Apfel,* lies wholly within the High German speech area and is customarily used to subdivide it into Middle German (*Appel*) and Upper German (*Apfel*) areas. Line 6, which indicates the shift of *Pund* 'pound' to *Pfund,* follows much the same course as does line 5 in the west, but it then runs north to join the *maken/machen* line; it is customarily used to distinguish West Middle German (*Appel, Pund*) from East Middle German (*Appel, Fund*—the latter being more common than Upper German *Pfund*).

As indicated above, Old High German grew out of the South Germanic branch, but both Old Saxon and Old Netherlandic also can be regarded as forms of South Germanic, albeit with differing ties to North Sea Germanic. Only after the consonant shift is there justification in speaking of a "(High) German" language distinct from the rest of South Germanic. The spread of early loans from Latin throughout all Germanic languages indicates that, at the time of their borrowing, a continuum of mutually intelligible dialects must still have existed, though differences between widely separated points in the continuum may have been significant. The modern German forms of these loans show that they were borrowed before the consonant shift, since they show its effects. Examples include Latin *pondō,* English *pound,* but German *Pfund;* Latin *piper,* English *pepper,* but German *Pfeffer;* Latin *tegula,* English *tile,* but German *Ziegel;* Latin (*via*) *strāta* 'paved (road),' English *street,* but German *Strasse;* Latin *catillus,* English *kettle,* but German *Kessel;* and Latin *coquus,* English *cook,* but German *Koch.*

Toward the end of the 4th century the great migrations of Germanic tribes resulted in an expansion of the Germanic-speaking territory. Angles, Saxons, and Jutes crossed the Channel to England; Franks moved southwest into northern France and southwestern central Germany; and Alemanni, Bavarians, and Langobardi moved south into southern Germany, Switzerland, Austria, and northern Italy. At the same time the area east of the Elbe and Saale rivers was largely vacated by Germanic speakers, and Slavic speakers moved in.

Old period (*c.* 750–1050)

In the southern area settled by Franks, Alemanni, and Bavarians, the first Old High German written records began to appear during the 2nd half of the 8th century. Their language is best described as a collection of monastery dialects; there is a certain uniformity in the writings of any given monastery but little for the area as a whole. The first documents are translations into German of Latin word lists. Later documents include prose translations of writings by St. Isidore of Seville (made *c.* 800) and by the Syrian Tatian (*c.* 830), as well as Otfrid's *Evangelienbuch* (*c.* 870), which introduced a new verse form with end rhyme. This literature reached its highest point in the able translations and interpretations of the Swiss scholar Notker Labeo (d. 1022). From the north (the Old Low German or Old Saxon speech area), the most extensive documents preserved are *Heliand* (*c.* 830), a life of Christ in alliterative verse, and a fragment of a poem based on the book of Genesis.

In this period there were many borrowings from Latin, nearly all connected with Christianization of the Germans. Because these borrowings were made after the consonant shift, they do not show its effects. Examples include *predigōn* (modern German *predigen* 'to preach'), from Latin *praedicāre; tempal* (modern German *Tempel* 'temple'), from Latin *templum;* and *spiagal* (modern German *Spiegel* 'mirror'), from Latin *speculum.* On the other hand, borrowings of this period reflect sound changes that had occurred in popular Latin, such as the change of Latin *c* before *e* from a *k* sound to *ts* in *cella* 'cell' and *crucem* 'cross,' borrowed into Old High German as *zella, krūzi,* modern German *Zelle, Kreuz* (the letter *z* in the German and Old High German examples represents the sound of *ts*); or the change of Latin medial *-b-* to *-v-* in *tabula* 'table,' borrowed into Old High German as *tavala,* modern German *Tafel.*

Middle period (c. 1050–1350)

Several developments justify the usual assumption of a new period, the language of which is called Middle High German, beginning in roughly 1050. First, there were changes in the language itself, among which were the unvoicing of final *b, d,* and *g* (compare Old High German *grab* 'grave,' *rad* 'wheel,' and *tag* 'day' with Middle High German *grap, rat,* and *tac;* in modern German these words are again spelled *Grab, Rad,* and *Tag* but are pronounced with final /p/, /t/, and /k/) and the reduction of the vowels of unstressed syllables to a /ə/, usually spelled *e* (*e.g.,* in the plural of the word for 'day,' the Old High German nominative-accusative form was *taga,* the genitive was *tago,* and the dative was *tagun,* but for these Middle High German had *tage, tage,* and *tagen,* respectively, and modern German has *Tage, Tage,* and *Tagen*). Second, there were great changes in the geographic area in which German was spoken. In the west the Franks of northern France had become romanized, and the French-German language border had assumed approximately its present location; in the east, on the other hand, German began to spread into Slavic territory, a process that was to continue for many centuries and to be reversed only at the end of World War II. Third, writing became independent of the monasteries, and the number of written documents soon increased greatly in both north and south. In the south, especially, a remarkable literature developed that included the courtly epic and *Minnesang.* There is clear evidence of a trend toward a standard Middle High German literary language, though it seems to have had no influence on ordinary speech. Because this literature was based largely on French models, many French words were borrowed into German.

Early modern period (c. 1350–1650)

Four events—the growth of trade, the rise of a middle class, the invention of the printing press, and the Reformation—had great influence on the development of the language. In the north, because of the prosperity of the Hanseatic League, a standard Low German written language began to develop, though it never reached full growth and probably had little influence on everyday speech. In the south the dialects that had arisen in the recently settled East Middle German area were relatively uniform and contained elements from both West Middle German and Upper German. Gradually these East Middle German dialects came to be used as the official languages of the chancelleries of the area, including that of Saxony, and on the latter—the East Middle German dialect of Saxony—Martin Luther based the language of his widely read Bible translation (1522–34). The growth of this type of German, which developed gradually into modern Standard German, was aided by the printers' desire to make their books appeal to the widest possible audience.

Three striking vowel changes are characteristic of this period. In the southeast, as early as the 12th century, the long vowels *ī, ū,* and *ǖ* (IPA *y*) came to be diphthongized to *ei, ou,* and *öü* (IPA *øy*); this feature is known as the New High German diphthongization. By the 15th century these new diphthongs had spread to East Middle German, and in the standard language they merged with the old diphthongs *ei, ou,* and *öü.* Examples include Middle High German *mîn* 'my,' *hûs* 'house,' and *hiuser* 'houses' with the monophthongs *ī, ū,* and *ǖ,* in contrast to *ein* 'a,' *troum* 'dream,' and *tröume* 'dreams' with the diphthongs *ei, ou,* and *öü,* but modern Standard German *mein, Haus,* and *Häuser* appear with the same diphthongs (*ai, au,* and *oi*) as *ein, Traum,* and *Träume.*

By a specifically Middle German development, the diphthongs *iə, uə,* and *üə,* still preserved in the southern dialects, were monophthongized to long *ī, ū,* and *ǖ;* this is the New High German monophthongization. Examples include Middle High German *tief* 'deep,' *vuoz* 'foot,' and *vüeze* 'feet' with the diphthongs *iə, uə,* and *üə,* contrasted with modern Standard German *tief, Fuss,* and *Füsse* with the monophthongs *ī, ū,* and *ǖ.* Short vowels remained short in closed syllables before long consonants but were lengthened in open syllables before a short consonant plus an unstressed vowel.

Modern period (since 1650)

The outstanding developments of the modern period have been the increasing standardization of High German and its increasing acceptance as the supradialectal form of the language. In writing, it is almost the only form used (except for limited printings of dialect literature); in speech, it is the first or second language of virtually the entire population.

Although Standard German is clearly based on the East Middle German dialects, it is not identical with any one of them; it has accepted and standardized many forms from other areas, notably the Upper German sound *pf* (*Pfund, Apfel*) and also large numbers of individual words in the forms of other dialect areas. Because it is the only type of German taught in schools, its spoken form is based to a large extent on its written form; and the spoken form that carries the greatest prestige (that of stage, screen, radio, and so on) uses a largely Low German pronunciation of this written form. As a result, the spoken form of modern Standard German has often been aptly described as "High German with Low German sounds."

Characteristics of modern Standard German. German has the following consonants, given here in phonetic symbols because the spelling often varies: stops, *p, b, t, d, k, g;* fricatives, *f, v, ç~x;* sibilants, *s, z, š, ž;* nasals, *m, n, ŋ;* liquids, *l, r;* glides, *h, j.* German *ç~x,* spelled *ch,* is the voiceless velar fricative *x* after *a, ā, o, ō, u, ū,* and *au* but is the voiceless palatal fricative *ç* in other phonetic environments. The German sound *ž* occurs only in loanwords from other languages.

In the orthography, German *w* always indicates a *v* sound symbolized /v/; German *v* spells an *f* sound in native words but a *v* sound in loanwords. German *sp* and *st* spell the sounds *sp* and *st* in most positions, but they spell *šp* /shp/ and *št* /sht/ at the beginnings of words or word stems. In other positions the *š* (*sh*) sound is spelled *sch—e.g., Schiff* 'ship.' German *z* always indicates /ts/. The spelling *tz* marks a preceding vowel as short, and the spelling *z* marks it as long.

Voiced *b, d, g, v,* and *z* do not occur at the ends of words, at the ends of parts of compound words, before suffixes beginning with a consonant, or before endings in *s* or *t.* In these positions they are replaced in pronunciation (though not in spelling) by the corresponding voiceless consonants, namely /p/, /t/, /k/, /f/, and /s/. For example, the *g* in *Tage* 'days' is pronounced as English *g,* and the *g* in *Tag* 'day' is pronounced as English *k.*

The German vowel system is given in Table 22 in phonetic symbols.

Table 22: Vowel System of German

short and lax vowels	long and tense vowels	diphthongs	unstressed vowel
i ü u	ī ǖ ū		
e ö o	ē ȫ ō	oi	ə
a	ā	ai au	

Though the spelling does not always indicate the difference between short and long vowels, the following devices are used more or less consistently: (1) A vowel is always short if followed by a double consonant letter—*e.g., still* 'still,' *wenn* 'if,' *Rasse* 'race,' *offen* 'open,' *Hütte* 'hut'—in contrast to the long vowels of *Stil* 'style,' *wen* 'whom,' *Glas* 'glass,' *Ofen* 'oven,' *Hüte* 'hats.' (2) A vowel is always long if followed by an (unpronounced) *h—e.g., ihnen* 'to them,' *stehlen* 'to steal,' *Kahn* 'barge,' *wohnen* 'to dwell,' *Ruhm* 'fame'—in contrast to the short vowels of *innen* 'inside,' *stellen* 'to place,' *kann* 'can,' *Wonne* 'bliss,' *dumm* 'dumb.' (3) A vowel is always long if written double—*e.g., Beet* '(flower) bed,' *Staat* 'state,' *Boot* 'boat'—in contrast to the short vowels of *Bett* 'bed (for sleeping),' *Stadt* 'city,' *Gott* 'god'; *ie* counts as the doubled spelling of *i—e.g.,* long *i* (*ī*) in *Miete* 'rent' but short *i* (*i*) in *Mitte* 'middle.' (4) A vowel (except unstressed *e*) is always long when it stands at the end of a word.

The "plain" vowels—*a, o, u, ā, ō, ū, au*—often alternate with the "umlaut" vowels—*e, ö, ü, ē, ō, ū, oi,* respectively—as in the following examples with plain vowels in the singular but umlauted vowels in the plural: *Gast* 'guest,' *Gäste; Gott* 'god,' *Götter; Mutter* 'mother,' *Mütter.* As these examples show, the vowel sounds *e, ē,* and *oi* are spelled *ä, ä,* and *äu* when they are the umlaut of *a, ā,* and *au* sounds. *Gast–Gäste, Vater–Väter, Braut–Bräute.* Otherwise they are generally spelled *e, eh,* or *ee* (*beten* 'to pray,' *geht* 'goes,' *Beet* '[flower] bed'), and *eu* (*Leute* 'people').

The sound /ai/ is generally spelled *ei: Seite* 'side,' *nein* 'no,' though in a few words *ai: Saite* 'string (of an instrument),' *Kaiser* 'emperor.' The schwa sound, /ə/, pronounced as the unstressed *a* in English *sofa,* is spelled *e: beginnen/bəgínən/*'to begin,' *geredet/gərēdət/*'spoken.'

(Wi.G.M./A.F.Bu.)

Yiddish. Although there were some 11 million speakers of Yiddish before World War II, approximately half of them were killed in the Nazi Holocaust. There are several million Yiddish speakers today, including native speakers and those who use it as a second language. Most speakers live in the United States and Israel. They are served by literary journals and an active press, including a number of daily newspapers.

International nature of Yiddish

History. Yiddish, although Germanic, is not a typical Germanic language; it includes not only Germanic features but also elements from Romance, Hebrew-Aramaic, and Slavic languages. A cursory examination of the German component of Yiddish indicates that no Yiddish dialect stands in a one-to-one relationship to any German dialect. The language had its beginnings in the 10th century when Jews from northern France and northern Italy settled in the Rhineland. These early Jewish settlements were dislocated by the Crusades and later by the persecutions that followed in the wake of the Black Death. The subsequent move to Slavic territory had enormous influence on the development of the language.

Onomastic evidence (evidence from recorded proper names) for Yiddish is known from 1096, and glosses in biblical commentaries are several decades older. The earliest known connected text is a rhymed couplet inscribed in a Hebrew holiday prayer book from Worms that bears the date 1272–73. The earliest extensive manuscript, known as the Cambridge Yiddish Codex, is explicitly dated Nov. 9, 1382. It excites the interest of Germanicists for its version of "Dukus Horant" (a poem from the Hildesage of the Kudrun [Gudrun] epic known from the Ambras Manuscript copied by Hans Ried, 1502/04–16), which antedates the earliest extant manuscript of the Hildesage by at least 130 years. The documentary history of Yiddish is unbroken thereafter to the present day. Unique evidence for spoken Yiddish is incorporated in an extensive body of rabbinical Responsa (published rabbinical opinions on matters of religious law) beginning in the 15th century. Testimony before the rabbinical court, which was recorded verbatim, provides unusual insight into the colloquial language.

Scholars divide the history of Yiddish into four periods: Earliest Yiddish, to 1250; Old Yiddish, 1250–1500; Middle Yiddish, 1500–1750; and Modern Yiddish, 1750 to the present. The earliest literary tradition had a Western Yiddish dialectal base; writing in this literary dialect continued into the Modern Yiddish period long after the major population centres had shifted to the east. The establishment of the modern literary language on an Eastern Yiddish base occurred only in the early 19th century. At the same time a new style in the language of Yiddish Bible translation emerged, free from the constraints of the original Hebrew syntax and of the stricture against the use of Yiddish words of Hebrew-Aramaic origin in translating from Hebrew. The continuous contact of Yiddish speakers with Hebrew-Aramaic texts and, in the European language area, with one or another Germanic or Slavic language have been important factors in the development of the language.

Characteristics. Because of the conditions under which Yiddish developed (*i.e.,* the numerous contacts it has had with other languages), it is of great interest to scholars.

Yiddish uses all the letters of the Hebrew alphabet, including traditional word-final variants, which in 1961 were reintroduced into the orthography of Russian Yiddish. Several letters occur only in words of Hebrew-Aramaic origin, which retain their traditional spelling in many countries.

The vowel system of Standard Yiddish consists of the simple vowels *i, e, a, o,* and *u* and the diphthongs *ej, aj,* and *oj.* Under Slavic influence a palatal series of consonants has emerged. The Yiddish *x* corresponding to German *ch* unlike German has no palatal variant, the /ng/ sound is simply a positional variant of *n,* there is no glottal stop (a sound made by closure of the vocal cords), and word-final voicing is distinctive (phonemic; *i.e.,* it carries a change in meaning). Words of Hebrew-Aramaic and Slavic origin have introduced a rich variety of consonant clusters that do not appear in German. Intonation contours, apparently related to the chant with which the Talmud is studied, convey syntactic-semantic distinctions independently.

Yiddish grammar

Case inflections, preserved only in the singular, appear in noun modifiers but only rarely in nouns themselves. The dative and accusative cases have merged in the masculine; the nominative and accusative cases have merged in the feminine and neuter. All prepositions govern the dative case. The system of forming noun plurals, basically of German origin, is enriched by word elements of Hebrew origin. Many nouns differ from their German cognates in both gender and plural form. A well-developed system of diminution uses word elements largely of German origin but on a Slavic grammatical model. A semantically significant distinction between inflected and uninflected predicate adjectives has emerged, while the difference between weak and strong adjectives, a characteristic of other Germanic languages, has effectively disappeared. The verb is inflected only in the present indicative. Other tenses and moods are expressed by means of auxiliary words. In normal word order the inflected verb immediately follows the subject; any remaining part of the verb phrase occurs as close to the inflected verb as possible. The special word order of the German subordinate clause is unknown, and verb-initial constructions generally express consecutiveness rather than interrogation.

In the Yiddish vocabulary, words and word elements borrowed from a number of different languages occur together and often combine freely in a manner unfamiliar to the languages from which they derive. Furthermore, when words borrowed from different languages are partially alike, one of them may be analyzed and inflected in terms historically appropriate to the other, thereby yielding blends of complex etymology. In addition, a highly productive system of prefixing yields verbs that are German in form but derive their meanings from an underlying Slavic model.

Dialects. The basic dialectal division is between Western Yiddish, which occurs largely within the German language area, and Eastern Yiddish in the Slavic-speaking areas. Eastern Yiddish is traditionally subdivided into Northeastern Yiddish and Southern Yiddish, the latter consisting of Central Yiddish and Southeastern Yiddish. The phonological criteria on which this division is based are typically reflected in the variants of the phrase 'to buy meat': Western Yiddish *kāfn flāš,* Central Yiddish *kojfn flajš,* Southeastern Yiddish *kojfn flejš,* Northeastern Yiddish *kejfn flejš.* Other phonological and many lexical differences reinforce the distinctness of Western Yiddish. In the east, Central Yiddish is further distinguished by a full set of contrasts in vowel length, while the varieties of Southeastern Yiddish have made changes in vowel quality that have led to the types *hont* 'hand,' *huz* 'house,' and *rign* 'rain.' Northeastern Yiddish is characterized by the loss of the neuter gender. Standard Yiddish adheres more closely to Northeastern Yiddish in its sound system and more closely to Southern Yiddish in its grammatical patterns. (M.I.H.)

NORTH GERMANIC (THE SCANDINAVIAN LANGUAGES)

History of Old Scandinavian. About 125 inscriptions dated from AD 200 to 600, carved in the older runic alphabet (futhark), are chronologically and linguistically the oldest evidence of any Germanic language. Most are from

Runes

Scandinavia, but enough have been found in southeastern Europe to suggest that the use of runes was also familiar to other Germanic tribes. Most inscriptions are brief, marking ownership or manufacture, as on the Gallehus Horns (Denmark; *c.* AD 400): *Ek Hlewagastiz Holtijaz horna tawido* 'I, Hlewagastiz, son of Holti, made [this] horn.' A number of inscriptions are memorials to the dead, while others are magical in content. The earliest were carved on loose wooden or metal objects, while later ones were also chiseled in stone. Further information about the language is derived from names and loanwords in foreign texts, from place-names, and from comparative reconstruction based on related languages and later dialects.

The inscriptions retain the unstressed vowels that were descended from Germanic and Indo-European but were lost in the later Germanic languages—*e.g.,* the *i*'s in *Hlewagastiz* and *tawido* (Old Norse would have been **Hlēgestr* and **táða*) or the *a*'s in *Hlewagastiz, Holtijaz,* and *horna* (Old Norse **Høltir, horn*). The scantiness of the material (fewer than 300 words) makes it impossible to be sure of the relationship of this language to Germanic and its daughter languages. It is known as Proto-Scandinavian, or Ancient Scandinavian, but shows few distinctively North Germanic features. The earliest inscriptions may reflect a stage, sometimes called Northwest Germanic, prior to the splitting of North and West Germanic (but after the separation of Gothic). Only after the departure of the Angles and Jutes for England and the establishment of the Eider River in southern Jutland as a border between Scandinavians and Germans is it reasonable to speak of a clearly Scandinavian or North Germanic dialect.

The emergence of Old Scandinavian, 600–1500. Inscriptions from the latter part of the Ancient period show North Germanic as a distinct dialect. Information about the earliest stages of the Old Scandinavian period is also derived from runic inscriptions, which became more abundant after the creation of the short runic futhark about AD 800. The expansion of Nordic peoples in the Viking Age (*c.* 750–1050) led to the establishment of Scandinavian speech in Iceland, Greenland, the Faroe Islands, the Shetland Islands, the Orkney Islands, the Hebrides, and the Isle of Man, as well as parts of Ireland, Scotland, England, France (Normandy), and Russia. Scandinavian languages later disappeared in all these territories except the Faroes and Iceland through absorption or extinction of the Scandinavian-speaking population.

Viking expansion

During the period of expansion, all Scandinavians could communicate without difficulty and thought of their language as one (sometimes called "Danish" in opposition to "German"), but the differing orientations of the various kingdoms in the Viking Age led to a number of dialectal differences. It is possible to distinguish a more conservative West Scandinavian area (Norway and its colonies, especially Iceland) from a more innovative East Scandinavian (Denmark and Sweden). An example of a linguistic difference setting off the eastern dialect area is the monophthongization of the Old Scandinavian diphthongs *ei, au,* and *øy* to *ē* and *ø* (*e.g., steinn* 'stone' became *stēn, lauss* 'loose' became *løs,* and *høyra* 'hear' became *hōra*). The diphthongs remained on the island of Gotland and in most North Swedish dialects, however, while they were lost in some East Norwegian dialects. The pronoun *ek* 'I' became *jak* in East Scandinavian (modern Danish *jeg,* Swedish *jag*) but remained *ek* in West Scandinavian (New Norwegian and Faroese *eg,* Icelandic *ég*); in East Norwegian it later became *jak* (dialects *je, jæ,* Dano-Norwegian *jeg*) but remained *ek* (dialects *a, æ*) in Jutland.

The advent of Christianity. The establishment of the Roman Catholic church during the 10th and 11th centuries had considerable linguistic significance. It helped to consolidate the existing kingdoms, brought the North into the sphere of classical and medieval European culture, and introduced the writing on parchment of Latin letters. Runic writing continued in use for epigraphic purposes and for general information (several thousand inscriptions are extant, from 11th-century Sweden, especially, and also all the way from Russia to Greenland). For more sustained literary efforts, the Latin alphabet was used—at first only for Latin writings but soon for native writings as well. The oldest preserved manuscripts date from approximately 1150 in Norway and Iceland and approximately 1250 in Denmark and Sweden. The first important works to be written down were the previously oral laws; these were followed by translations of Latin and French works, among them sermons, saints' legends, epics, and romances. Some of these may have stimulated the extraordinary flowering of native literature, especially in Iceland. One can hardly speak of distinct languages in this period, although it is customary to distinguish Old Icelandic, Old Norwegian, Old Swedish, Old Danish, and Old Gutnish (or Guthnic, spoken in Gotland) on the basis of quite minor differences in the writing traditions. Some of these were merely scribal habits resulting from local usage, but others did reflect the growing separation of the kingdoms and the centralization within each. Literary Old Icelandic is often presented in a normalized textbook form and (together with Old Norwegian) is referred to as Old Norse.

Roman influence

Culture words like *caupō* 'merchant' (giving Old Norse *kaupa* 'buy') and *vinum* 'wine' (Old Norse *vín*) had been filtering into the North from the Roman Empire for a long time. But the first great wave of such words came from the medieval church and its translations, often with the other Germanic languages as intermediaries because the first missionaries were English and German. Some religious terms were borrowed from other Germanic languages; among these are Old Norse *helviti* 'hell' from Old Saxon *helliwiti* or Old English *hellewite,* and Old Norse *sál* 'soul' from Old English *sāwol.* East Scandinavian borrowed the Old Saxon word *siala,* from which come later Danish *sjæl* and Swedish *själ.* In the secular field the most profound influence on Scandinavian was that exerted by Middle Low German because of the commercial dominance of the Hanseatic League and the political influence of the North German states on the royal houses of Denmark and Sweden between 1250 and 1450. The major commercial cities of Scandinavia had large Low German-speaking populations, and the wide use of their language resulted in a stock of loanwords and grammatical formatives comparable in extent to that which French left behind in English after the Norman Conquest.

Reformation and Renaissance. The many local dialects that exist today developed in the late Middle Ages, when the bulk of the population was rural and tied to its local village or parish, with few opportunities to travel. The people of the cities developed new forms of urban speech, coloured by surrounding rural dialects, by foreign contacts, and by the written languages. The chanceries in which documents of government were produced began to be influential in shaping written norms that were no longer local but nationwide. The Reformation came from Germany and with it Martin Luther's High German translation of the Bible, which was quickly translated into Swedish (1541), Danish (1550), and Icelandic (1584). That it was not translated into Norwegian was one of the major reasons that no separate Norwegian literary language arose. Literary Old Norwegian went out of use, and until the 19th century there was no distinct written Norwegian. Instead a Norwegian variety of Danish developed and became the basis of Dano-Norwegian Bokmål. With the invention of printing and the growth of literacy, all speakers of Scandinavian dialects gradually learned to read (and eventually write) the new standard languages.

Languages of the subgroup. The six standard languages of today, in the order of their emergence as languages of culture and prestige, are Danish, Swedish, Icelandic, Faroese, New Norwegian (Nynorsk), and Dano-Norwegian.

Danish. The norms of the first printed books in Danish continued the norm of the royal chancery in Copenhagen, which was not based on any particular dialect and probably reflected a state of the language closer to that of 1350 than of 1550. Because of the influence of the written language, many speech forms used even by the aristocracy at that time were eliminated or branded as vulgar. Danish is clearly the Scandinavian language that has undergone the greatest amount of change away from the Old Scandinavian norm. In the 18th century a mildly puristic reform led to the replacement of many French loans by their native equivalents (*e.g., imagination* was replaced by *indbildning;*

compare German *Einbildung*), and, in the 18th and 19th centuries, Danish became the vehicle of a classical literature. There are regional differences in the cultivated speech norm, but upper-class Copenhagen speech probably has the highest prestige. A spelling reform in 1948 eliminated the capitalization of nouns and introduced the letter *å* for *aa*, thereby making the spelling more similar to that of Norwegian and Swedish. Danish is spoken by most of the more than five million inhabitants of Denmark and in a few communities south of the German border; it is an official language of the Faroe Islands and Greenland and is the first foreign language taught in Iceland.

Spelling reform

Swedish. Before the Swedish revolt of Gustav I Vasa in 1525, Danish influence on the Swedish language had been strong; the new government, however, made vigorous efforts to eliminate this. The written norm was based on one that had developed in the manuscripts of central Sweden, extending from the Vadstena monastery in eastern Götaland to Stockholm and Uppsala. In comparison to the speech of the area, many of its features were conservative (*e.g.,* silent *-t* and *-d* in words like *huset* 'the house' and *kastad* 'thrown'). The written language was cultivated energetically as a symbol of national strength, and in 1786 the Swedish Academy was established by King Gustav III. The language spread at the expense of Danish by the conquest of southern and western provinces in the 17th century. After Sweden ceded Finland to Russia in 1809, the role of Swedish was gradually reduced in that country. From 1917 to 1995 Swedish was an official language of Finland; after mid-1995 Finland had no official languages, including Finnish. Both Finnish and Swedish are national languages. Swedish is spoken by about 90 percent of the population of Sweden (some eight million), and Swedish-language literature is rich and distinguished.

Icelandic. Important factors in the survival of Icelandic during the period of Danish rule were its continued use for literary purposes, the geographic remoteness of Iceland, a scattered population, and the great linguistic differences between Danish and Icelandic. While the Scandinavian languages in continental Europe were losing inflection, Icelandic preserved Old Scandinavian grammar almost intact. The native Bible became a basis for the further development of Icelandic. Nevertheless, the circumstances of the language were highly restricted until self-government developed in the 19th century and Icelandic was rediscovered by Scandinavian scholars. A firm orthography along etymological lines was gradually established, and the policy of not adopting foreign words was confirmed, so that Icelandic today appears strikingly different from the other Scandinavian languages. Icelandic is spoken by the entire population of the country, some 290,000 people.

Faroese. Prior to modern times literary activity in the Faroe Islands was minimal, but the local dialects continued to develop, though Danish was the official language. The Danish language scholar Rasmus Rask, who wrote the first Faroese grammar (1811), described the language as a dialect of Icelandic, but it is actually an independent language, intermediate between West Norwegian and Icelandic and containing many Danish loanwords. Traditional dance ballads were written down after 1773 before the establishment in 1846 of an independent orthography. This orthography is etymologizing and unphonetic and gives Faroese a strong Icelandic appearance. The establishment of home rule in 1948 led to the introduction of Faroese as the primary language taught in the schools. The language is now spoken by some 48,000 people on the Faroe Islands.

Faroese spelling

Norwegian. Old Norwegian writing traditions gradually died out in the 15th century after the union of Norway with Denmark and the removal of the central government to Copenhagen. After independence was achieved in 1814, the linguistic union with Danish persisted, but educational problems due to the linguistic distance between Danish and spoken Norwegian and socio-political considerations, as well as the ideology of "national Romanticism," stimulated a search for a national standard language. In 1853 a young, self-taught linguist of rural stock, Ivar Aasen, constructed a language norm from the spoken dialects. This standard continued the Old Norwegian tradition and was meant to eventually replace Danish. After long research and experimentation, he presented this New Norwegian norm (called Landsmål, but now officially Nynorsk) in a grammar, a dictionary, and numerous literary texts. New Norwegian was officially recognized as a second national language in 1885. Today all Norwegians learn to read and write it, but only a fifth of the population use it as their primary written language. It has been cultivated by many excellent authors and has a quality of poetic earthiness that appeals even to nonusers. Its norm has changed considerably since Aasen's time in the direction of spoken East Norwegian or written Dano-Norwegian.

New Norwegian

Most Norwegian literature in the 19th century was written in a superficially Danish norm, but it was given Norwegian pronunciation and had many un-Danish words and constructions. The spoken norm was a compromise Dano-Norwegian that had grown up in the urban bourgeois environment. In the 1840s Knud Knudsen formulated a policy of gradual reform that would bring the written norm closer to that spoken norm and, thereby, create a distinctively Norwegian language without the radical disruption envisaged by the supporters of Aasen's New Norwegian. This solution was supported by most of the new writers in the powerful literary movement of the late 19th century. The official reforms of 1907, 1917, and 1938 broke with the Danish writing tradition and adopted native pronunciation and grammar as its normative base; the resultant language form was called Riksmål, later officially Bokmål. An effort aiming at amalgamating Dano-Norwegian and New Norwegian into one language (Samnorsk) has been abandoned as an official language policy. In its current form Dano-Norwegian is the predominant language of Norway's population of 4.6 million, except in western Norway and among the Sami minority in the North.

The teaching of the standard languages in the schools and the high levels of literacy have tended to spread the urban norms of speaking. Nevertheless, very diverse dialects, partially unintelligible to outsiders, are spoken in many rural communities; some of them are used occasionally for the writing down of local traditions or for giving local colour. Local dialects are used much more widely in Norway than in the other Scandinavian—and European—countries. It is not unusual for university professors, politicians, business executives, and other public figures to use their local dialects even when speaking in a professional capacity. Boundaries between dialect areas are gradual and do not always coincide with national borders, so that the following traditional divisions are somewhat arbitrary: in Denmark, West (Jutland), Central (Fyn, Sjælland), and East (Bornholm); in Sweden, South (especially Skåne), Götaland, Svealand, North (Norrland), Gotland, and East (Finland); in Norway, East (Lowland, Midland), Trönder (around Trondheim), North, and West. In the Faroese language there are minor dialectal differences between the southern and northern islands; minor dialectal differences occur in Icelandic as well, but there are no clearly defined regional dialects. In the larger cities social dialects range from the everyday speech of the working classes (often similar to nearby rural speech) to the more cultivated forms of middle- and upper-class speech, including the highly formal style of courts and legislatures. Speakers of Danish, Norwegian, and Swedish normally use their own languages in communicating with one another.

Linguistic characteristics. *Common and distinctly Scandinavian characteristics.* North Germanic differs from West Germanic (but not East Germanic) in having *ggj* and *ggv* for medial *jj* and *ww,* respectively (Old Norse *tveggja* 'two,' *hoggva* 'hew'), *-t* for *-e* in the second person singular of the strong preterite (Old Norse *namt* 'you took'; compare Old English *næme*), and a reflexive possessive *sin*.

North Germanic differs from East Germanic (but not West Germanic) in that original *ē* becomes *ā* (Old Norse *māni* 'moon') and original *z* becomes *r* (Old Norse *meiri* 'more'); furthermore, there is a new demonstrative pronoun *þessi* 'this' (Danish, Swedish, and Norwegian *denne*), back vowels are mutated to front vowels by the influence of *a* following *i* or *j* ("*i*-umlaut"—*a* and *ā* become *æ* and *ǣ*, *o* and *ō* become *ø* and *ø̄* [*ø* represents umlauted *o*], *u* and *ū* become *y* and *ȳ* [*y* represents umlauted *u*], *au*

A trait shared with West Germanic

becomes *ey* or *øy*), and the number of unstressed vowels is reduced to three (*a, i, u*).

North Germanic differs from both West Germanic and East Germanic in the following ways: rounding of unrounded vowels by following *u* or *w* ("*u*-umlaut"—*a* and *ā* become *ǫ* and *ǭ* [*ǫ* represents a low back rounded vowel], *e* becomes *ø*, *i* becomes *y*, *ei* becomes *ey* or *øy*); loss of initial *j* and of *w* before rounded vowels (Old Norse *ár* 'year,' *ungr* 'young,' *orð* 'word'); loss of final nasals (Old Norse *frá* 'from,' and generally in infinitives: Old Norse *fara* 'fare, go'; compare Old English *faran*, German *fahren*); diphthongization (the creation of a gliding monosyllabic speech sound) of short *e* to *ja* or *jǫ* (Old Norse *jafn* 'even,' *jǫrd* 'earth'). It has new pronouns for the third person singular (Old Norse *hann* 'he,' *hon* 'she'); attaches the reflexive pronoun (*sik*) to the verb to make a new mediopassive in *-sk, -st,* or *-s* (*finna sik* 'find oneself' became Old Norse *finnast* 'be found, exist,' Danish *findes*); attaches the demonstrative *inn* 'that' to nouns as a definite article (Old Norse *fótrinn* 'the foot,' Norwegian and Swedish *foten*, Danish *foden*), except in West Jutland (possibly a later development); and uses *-t* as marker of the neuter in pronouns and adjectives (Old Norse *stórt* 'big' from *stór-*).

Furthermore, North Germanic employed *es* (which changed to *er*) and later *sum* as an indeclinable relative pronoun. It also lost some Germanic prefixes such as *ga-* (German *ge-*) and contains a considerable number of words such as *hestr* 'horse,' *fær* or *fár* 'sheep,' *gríss* 'pig,' *gólf* 'floor,' and *ostr* 'cheese' that do not occur in East or West Germanic.

Orthography. The five basic vowel symbols of the Latin alphabet are supplemented by a number of special symbols that are used mostly to represent umlauted vowels: thus, there is *y* (pronounced as German *ü*), *æ* (used in Danish, Norwegian, Icelandic, and Faroese) and the corresponding *ä* (used in Swedish), *ø* (in Danish, Norwegian, and Faroese) and the corresponding *ö* (in Swedish and Icelandic), and *å* (also written *aa*, used in Danish, Swedish, and Norwegian).

Their present-day values are not identical; Icelandic *æ* is pronounced as the diphthong sound *ai* (as the *i* in English *ice*). Icelandic and Faroese also use accents on vowels that were long in Old Norse but are now mostly diphthongs (*á, é, í, ó, ú,* and *ý*). The consonant symbols are the usual Latin ones, except that *þ* (thorn) and *ð* (eth) are used in Icelandic for voiceless and voiced *th* (*ð* in Faroese has a different value). Loanwords containing the letters *c, q, w, x,* and *z* have generally been naturalized by substituting, respectively, *k* or *s, kv, v, ks,* and *s* (*e.g., kontakt* 'contact' but Norwegian *sigar* 'cigar' versus Danish and Swedish *cigar*).

Stress in North Germanic

Phonology. Stress is placed on the first syllable in native words, with sporadic exceptions for compounds. Stress on a later syllable reflects borrowing from other languages, except in Icelandic, which has stress on the first syllable of all words. (The latter is also is true of East Norwegian dialects.)

Pitch is usually high on the stressed syllable, falling at the end of a statement, rising for a yes-no question. An exception is East Norwegian and some Swedish dialects, in which the stressed syllable is low and the pitch is often rising at the end of statements. In most of Norway and Sweden and in scattered Danish dialects, there is a special word tone, by which old monosyllables have one kind of pitch while old polysyllables have another. The first pitch type is usually high or low pitch on the stressed syllable, like that in other Germanic languages, while the second is more complex and varies from region to region. In Danish the tones have been replaced by glottalization in instances in which Norwegian and Swedish have the first type.

In stressed syllables either the vowel or the following consonant is long (except in Danish). A short vowel also may be followed by a consonant cluster, but a long vowel may never be followed by a long consonant. Unstressed syllables may have a short vowel followed by a short (or no) consonant. In Danish this latter pattern is permitted also in stressed syllables.

The Old Scandinavian vowel system contained nine vowels, each of which could be long, short, or nasalized: front unround (*i, e, æ*), front round (*y, ø*), back round (*u, o, ǫ*), and back unround (*a*). There were three falling diphthongs (*ei, au, øy*). While most of these are still present in some dialects, there have been many changes. The nasalized vowels disappeared, though they were still present in Icelandic about 1150. Diphthongs became long vowels in Danish and Swedish in the 10th century. Short low umlauted vowels coalesced with neighbouring vowels (*æ* became *e* and *ǫ* became *o*, or *ö* in Icelandic). Long *ā* (Old Norse *á*) was rounded to *å* (pronunciation similar to the *o* in English *order*; in Icelandic and West Norwegian, pronunciation is like the *ow* in English *now*). In Norwegian and Swedish the rounded vowels were shifted upward and forward, giving "overrounded" *o* and *u* that resemble *u* and *y*, respectively. The unstressed vowels *a, i,* and *u* have remained in Icelandic and Faroese but have been partially merged in New Norwegian and Swedish (written *a, e, o*), completely merged as *ə* (the schwa sound, as *a* in English *sofa*) in Danish and Dano-Norwegian, and lost in Jutland and Trønder dialects. High round vowels (*y, ȳ, øy*) have been merged with the unround vowels in Icelandic and Faroese (and in scattered dialects elsewhere) but are still distinguished in writing. Long vowels have been diphthongized not only in many dialects (*e.g.,* Jutland, Skåne, and West Norwegian) but also in standard Icelandic and Faroese (Icelandic *é*, pronounced /je/, *ó* /ou/, *á* /au/, *æ* /ai/; Faroese *í* /ui/, *æ* /æa/, and so on). (Symbols in virgules are phonetic symbols designating actual pronunciation.) A quantity shift took place in the late Middle Ages, in which short vowels were lengthened before single consonants and long vowels were shortened before clusters, sometimes with qualitative changes that affected different dialects differently; thus, in Swedish *veta* 'know' *i* became *e* (though all the other North Germanic languages have *i*).

The Old Scandinavian consonant system

The Old Scandinavian consonant system contained voiceless stops *p, t, k;* voiced stops *b, d, g;* voiceless-voiced spirants *f/v, þ/ð, x/g;* nasals *m, n;* a sibilant *s;* liquids *l, r;* and glides *w, j.* The chief changes were as follows: Short voiceless stops became voiced after vowels in Danish and neighbouring dialects, and then they partially opened to become spirants or glides (*tapa* became *tabe* 'lose,' *ūt* became *ud* 'out,' *kakur* became *kager* 'cakes'). Velar stops *k, g,* and *sk* were palatalized before front vowels to merge with *kj, gj,* and *skj,* as still occurs in Icelandic (and Jutland dialect); in Faroese, Norwegian, Swedish, and many Danish dialects, these were fronted to *tj, dj,* and *stj* or even opened to spirants *ç, j,* and *š,* while in Danish they reverted to *k, g,* and *sk.* Voiced *f* merged with *w* to become *v,* though it is still written *f* in Icelandic; in Danish both *f* and *w* have become pronounced as *w* after vowels. Voiceless *þ* became *t* (occasionally *h* in Faroese) and voiced *þ* /ð/ became *d,* except in Icelandic. Voiceless *x* became *h* initially before vowels but was lost elsewhere; voiced *x* /g/ became *g,* except in Icelandic (in Danish it has become either /j/ or /w/ after vowels). The *r* sound was assimilated to following dental sounds (*l, n, s, t, d*) to make a series of retroflex consonants (*ḷ, ṇ, ṣ, ṭ, ḍ,* pronounced with the tip of the tongue curled up toward the hard palate) in many Swedish and Norwegian dialects, including those of Oslo and Stockholm. In western Sweden and eastern and central Norway, an original *l* in certain environments and the combination *rð* both developed into a new sound defined as a retroflex flap. During the past few centuries the *r* sound has become a uvular *r* /ʀ/ in Danish, southern Swedish, and southwestern Norwegian (including that spoken in the city of Bergen). The uvular *r* is still expanding its territory in Norway and Sweden.

Morphology. Old Scandinavian had a declensional system with four cases (nominative, accusative, dative, genitive) and two numbers. The actual form of the inflections depended on the stem class of the noun or the adjective. Verbs were inflected for tense and mood, person and number. This system is preserved in Icelandic. In Faroese, the declensions have been simplified, and only three cases are now used in speech (the genitive having been replaced by prepositional phrases or compounds). In the remaining languages only personal pronouns now have a distinction between a nominative and a non-nominative (dependent) form (*e.g.*, Swedish *jag* 'I,' *mig* 'me'). Some conservative dialects in Norway and Sweden still retain a separate dative case for certain categories, however.

Major similarities

The present-day systems of Danish, Dano-Norwegian, New Norwegian, and Swedish are basically identical. Nouns have singular and plural forms, to which the definite article may be suffixed; the plural suffixes vary, reflecting earlier stem, gender, and umlaut classes. Adjectives have neuter singulars marked by *-t*, plurals marked by a vowel (*-e* or *-a*), and weak forms used after determiners, usually identical in form with the plurals; the comparatives are marked by *r* and superlatives by the cluster *st*. There are polite pronouns of address that are either identical with the second person plural (Swedish *ni*, Icelandic *þér*, Faroese *tygum*, and New Norwegian *de*) or the third person plural (Danish and Dano-Norwegian *De*). In Norway and Sweden the use of the polite form is now obsolete. In Icelandic and Faroese old duals have taken over the function of plurals (Icelandic *við* 'we,' *þið* 'you'; Faroese *vit* 'we,' *tit* 'you'). Each personal pronoun has a corresponding possessive pronoun, the third person being identical with the genitive of the pronoun and invariable. The possessive pronouns for the other persons and the reflexive *sin* are inflected for gender and number like most other pronouns and articles. Verbs inflect for tense only, with *-r* as the usual present marker (New Norwegian does not have an ending to indicate present tense in the strong verbs), while the preterites (past tenses) have stem-vowel ablaut changes in the strong verbs and a dental suffix in the weak verbs. Nonfinite forms of the verb have invariable suffixes (*-a* or *-e* for the infinitive, *-ande* or *-ende* for present participles, and *-at* or *-et* for perfect participles), except that Swedish and New Norwegian mark gender when the perfect participle is used adjectivally.

Major differences

New Norwegian, like Icelandic and Faroese, and, in part, Dano-Norwegian preserve masculine, feminine, and neuter genders; Danish and Swedish combine masculine and feminine into a common (nonneuter) gender. Swedish and New Norwegian (in part) preserve nonneuter plurals in *-ar*, *-er*, and *-or*, which merged as *-er* in Dano-Norwegian; in Danish these have become *-e*, while a new plural in *-er* has arisen, primarily for loanwords. The past tense of the largest class of weak verbs (Old Norse *-aði*) ends in *-a* in New Norwegian, *-et* or *-a* in Dano-Norwegian, *-ede* in Danish, and *-ade* in Swedish (usually pronounced /a/). In Norwegian and Swedish a new class of weak verbs with preterite ending *-dde* has arisen, including stems ending in *-d* or long vowels (Swedish *födde* 'bore,' *bodde* 'lived'). The present tense form of strong verbs is umlauted in New Norwegian (as in Icelandic and Faroese); it is monosyllabic in New Norwegian, has high or low pitch on the stressed syllable in Dano-Norwegian and Swedish, and has glottalization in Danish (New Norwegian *kjem*; Dano-Norwegian, Swedish *kommer* /kåmər/; Danish *kommer* /kåmˀər/. New Norwegian has *-st* in the mediopassive (like Icelandic and Faroese); Dano-Norwegian, Swedish, and Danish have *-s*. Besides a complex passive formed with an auxiliary, Swedish, Danish, Dano-Norwegian, and (to a limited degree) New Norwegian have developed an inflectional passive form in *-s* by the reduction of the old reflexive pronoun *sik*.

Syntax. The reduction of morphological complexity has been accompanied by the emergence of a more rigid order of sentence elements. Main clauses have the finite verb in second position. This can be preceded by almost any other sentence constituent; most often it is preceded by the subject. In yes-no questions the preverbal position is empty. In other questions it is occupied by the question word. When the subject does not precede the verb, it follows it. A nonfinite verb follows the subject but precedes the object and adverbials (except sentence adverbials and certain time adverbials, which may precede the nonfinite verb). In Icelandic, subordinate clauses have the same basic structure as main clauses; in the other languages the verb always follows the subject and any sentence adverbial. Complex verb phrases are formed with modal auxiliaries (*e.g.*, *kan* 'can') and infinitives or with the perfect auxiliaries *ha(ve)* 'have' and *få* 'get' (Icelandic *geta*) and the perfect participle. Instead of such durative aspect markers as the English progressive (*e.g.*, "is talking"), verbs indicating position are combined with the main verb (*e.g.*, Dano-Norwegian *han sitter* [*står, går, ligger*] *og prater* 'he is sitting [standing, walking, lying] and talking.'). Icelandic has special constructions for present and perfect aspects (*er að ganga* 'is going' or *er buinn að ganga*, literally, 'is through going').

Major differences in the Norwegian languages, Swedish, and Danish are few: (1) New Norwegian and Swedish use the nominative after a copula (*Det er eg/jag* 'It is I'), Dano-Norwegian and Danish, the accusative (*Det er meg/mig* 'It is me'). (2) A complex passive is formed either with Old Scandinavian *verða* (Swedish *varda*, New Norwegian *verta*) or Low German *bliven* (Danish *blive*, Dano-Norwegian *bli*) and the perfect participle. (3) A verbal particle precedes the object in Swedish (*Jag brände upp den/tidningen* 'I burned it up/[I burned up] the newspaper'), follows it in Danish (*Jeg brendte den/avisen opp*), while both orders are used in Norwegian, depending on the relative weight of the particle and the object (*Eg brende henne opp/Eg brende opp avisa*). (4) The reflexive pronoun *sin* is used with singular or plural subjects, except in Danish, in which it is used only with singular subjects. (5) A definite article is indicated by a form before the adjective and a suffix after the noun ("double definite"), except in Icelandic and Danish (*e.g.*, in Norwegian and Swedish *det store* [*stora*] *huset* 'the big house,' both *det* and *-et* in *huset* mean 'the,' in Danish the suffix *-et* is not used: *det store hus*). (6) A possessive may follow its noun in Icelandic, Faroese, and Norwegian but not in Danish or Swedish (Icelandic *hesturinn minn* 'my horse,' literally, 'horse mine,' Swedish *min häst* 'my horse'). (7) The numeral 'one' is used (in unstressed form) as an indefinite article (*i.e.*, as *a* and *an* are used in English), except in Icelandic, which has no indefinite article. (8) Swedish omits the auxiliary *hava* 'have' in subordinate clauses (*Huset jag sett* . . . 'The house I [have] seen . . . ').

Vocabulary. The everyday stock of Scandinavian words, including most of the high-frequency words, is Indo-European and Germanic in its core. Of the 200,000 or more entries in the large dictionaries of each language, the vast majority are either compounds and derivatives of the simpler words or else borrowings from other languages—mostly of a scientific and cultural nature. At the end of the 20th century, the chief source of loanwords in the North Germanic languages was English.

Borrowings

Icelandic preserved the creative powers of the older language by making it a policy not to accept new words in unassimilated form. Whenever possible, new compounds and derivatives have been created to avoid the borrowing of foreign terms. To some extent Faroese and New Norwegian have followed the same policy but without the degree of success that Icelandic has had. Danish, Swedish, and Dano-Norwegian have adopted numerous German words, along with their prefixes and suffixes—*e.g.*, Danish and Norwegian *betale* and Swedish *betala* 'pay' from Low German *betalen*.

The borrowings of Danish, Swedish, and Norwegian reflect the varied contacts discussed above. Their vocabulary consists of a native core, a German middle layer (with words like Danish *skrædder* 'tailor'; compare Icelandic and Faroese *klæðskeri*, literally 'cloth-cutter'), and an international outer layer (with words such as *psykologi* 'psychology'; compare Icelandic and Faroese *sálfræði*, literally 'soul science'). While there are some differences among the languages in the exact composition of these layers, there is also considerable agreement. Differences occur especially in words of local origin (slang, humour, endearments, abuse) and in borrowings of different origin—*e.g.*, Norwegian *etasje*/Swedish *våning*/Danish *sal* 'story' (in a hotel), from French *étage*, Middle Low German *woninge*, and Old Scandinavian *salr* (but with its meaning from North German *Saal*). (Ei.H./J.T.F.)

English language

English is widely spoken in all six continents and has had a strong effect in many regions in which it is not the principal language spoken. The portmanteau words Franglais, Russlish, and Japlish, for example, have been invented by resentful purists to describe the numerous expressions, in vogue among the young, resulting from the infiltration of English into French, Russian, and Japanese, respectively.

The widely diffused English-speaking community is fairly stable in the British Isles, North America, and Australasia. In Africa, the Indian subcontinent, and Southeast Asia its future remains uncertain and unpredictable. People who speak English fall into three groups: those who have inherited it as their native language; those who have acquired it as a second language within a society that is largely bilingual; and those who are driven by necessity to use it for some practical purpose—administrative, professional, or educational. One person in seven of the world's entire population now belongs to one of these three groups.

ORIGINS AND BASIC CHARACTERISTICS

Parent tongue

English belongs to the Indo-European family of languages and is therefore related to most other languages spoken in Europe and western Asia from Iceland to India. The parent tongue, called Proto-Indo-European, was spoken about 5,000 years ago by nomads believed to have roamed the southeast European plains. Germanic, one of the language groups descended from this ancestral speech, is usually divided by scholars into three regional groups: East (Burgundian, Vandal, and Gothic, all extinct), North (Icelandic, Faeroese, Norwegian, Swedish, Danish), and West (German, Netherlandic [Dutch and Flemish], Frisian, English). Though closely related to English, German remains far more conservative than English in its retention of a fairly elaborate system of inflections. Frisian, spoken by the inhabitants of the Dutch province of Friesland and the islands off the west coast of Schleswig, is the language most nearly related to Modern English. Icelandic, which has changed little over the last thousand years, is the living language most nearly resembling Old English in grammatical structure.

Modern English is analytic (*i.e.*, relatively uninflected), whereas Proto-Indo-European, the ancestral tongue of most of the modern European languages (*e.g.*, German, French, Russian, Greek), was synthetic, or inflected. During the course of thousands of years, English words have been slowly simplified from the inflected variable forms found in Sanskrit, Greek, Latin, Russian, and German, toward invariable forms, as in Chinese and Vietnamese. The German and Chinese words for "man" are exemplary. German has five forms: *Mann, Mannes, Manne, Männer, Männern.* Chinese has one form: *jen.* English stands in between, with four forms: man, man's, men, men's. In English only nouns, pronouns, and verbs are inflected. Adjectives have no inflections aside from the determiners "this, these" and "that, those." (The endings *-er, -est,* denoting degrees of comparison, are better regarded as noninflectional suffixes.) English is the only European language to employ uninflected adjectives; *e.g.*, "the tall man," "the tall woman," compared to Spanish *el hombre alto* and *la mujer alta.* As for verbs, if the Modern English word ride is compared with the corresponding words in Old English and Modern German, it will be found that English now has only five forms (ride, rides, rode, riding, ridden), whereas Old English *ridan* had 13, and Modern German *reiten* has 16 forms.

In addition to this simplicity of inflections, English has two other basic characteristics: flexibility of function and openness of vocabulary.

Loss of inflection

Flexibility of function has grown over the last five centuries as a consequence of the loss of inflections. Words formerly distinguished as nouns or verbs by differences in their forms are now often used as both nouns and verbs. One can speak, for example, of "planning a table" or "tabling a plan," "booking a place" or "placing a book," "lifting a thumb" or "thumbing a lift." In the other Indo-European languages, apart from rare exceptions in Scandinavian, nouns and verbs are never identical because of the necessity of separate noun and verb endings. In English, forms for traditional pronouns, adjectives, and adverbs can also function as nouns; adjectives and adverbs as verbs; and nouns, pronouns, and adverbs as adjectives. One speaks in English of the Frankfurt Book Fair, but in German one must add the suffix *-er* to the place-name and put attributive and noun together as a compound, Frankfurter Buchmesse. In French one has no choice but to construct a phrase involving the use of two prepositions: Foire du Livre de Francfort. In English it is now possible to employ a plural noun as adjunct (modifier), as in "wages board" and "sports editor"; or even a conjunctional group, as in "prices and incomes policy" and "parks and gardens committee."

Openness of vocabulary implies both free admission of words from other languages and the ready creation of compounds and derivatives. English adopts (without change) or adapts (with slight change) any word really needed to name some new object or to denote some new process. Like French, Spanish, and Russian, English frequently forms scientific terms from Classical Greek word elements.

English possesses a system of orthography that does not always accurately reflect the pronunciation of words; this is discussed below in the section *Orthography.*

CHARACTERISTICS OF MODERN ENGLISH

Phonology. British Received Pronunciation (RP), by definition, the usual speech of educated people living in London and southeastern England, is one of the many forms of standard speech. Other pronunciations, although not standard, are entirely acceptable in their own right on conversational levels.

The chief differences between British Received Pronunciation, as defined above, and a variety of American English, such as Inland Northern (the speech form of western New England and its derivatives, often popularly referred to as General American), are in the pronunciation of certain individual vowels and diphthongs. Inland Northern American vowels sometimes have semiconsonantal final glides (*i.e.*, sounds resembling initial *w*, for example, or initial *y*). Aside from the final glides, this American dialect shows four divergences from British English: (1) the words cod, box, dock, hot, and not are pronounced with a short (or half-long) low front sound as in British "bard" shortened (the terms front, back, low, and high refer to the position of the tongue); (2) words such as bud, but, cut, and rung are pronounced with a central vowel as in the unstressed final syllable of "sofa"; (3) before the fricative sounds *s, f,* and *θ* (the last of these is the *th* sound in "thin") the long low back vowel *a*, as in British "bath," is pronounced as a short front vowel *a*, as in British "bad"; (4) high back vowels following the alveolar sounds *t* and *d* and the nasal sound *n* in words such as tulips, dew, and news are pronounced without a glide as in British English; indeed, the words sound like the British "two lips," "do," and "nooze" in "snooze." (In several American dialects, however, these glides do occur.)

The consonants

The 24 consonant sounds comprise six stops (plosives): *p, b, t, d, k, g;* the fricatives *f, v, θ* (as in "thin"), *ð* (as in "then"), *s, z, ʃ* (as in "ship"), *ʒ* (as in "pleasure"), and *h;* two affricatives: *tʃ* (as in "church") and *dʒ* (as the *j* in "jam"); the nasals *m, n, ŋ* (the sound that occurs at the end of words such as "young"); the lateral *l;* the vibrant or retroflex *r;* and the semivowels *j* (often spelled *y*) and *w*. These remain fairly stable, but Inland Northern American differs from British English in two respects: (1) *r* following vowels is preserved in words such as "door," "flower," and "harmony," whereas it is lost in British; (2) *t* between vowels is voiced, so that "metal" and "matter" sound very much like British "medal" and "madder," although the pronunciation of this *t* is softer and less aspirated, or breathy, than the *d* of British English.

Like Russian, English is a strongly stressed language. Four degrees of stress may be differentiated: primary, secondary, tertiary, and weak, which may be indicated, respectively, by acute (´), circumflex (ˆ), and grave (`) accent marks and by the breve (˘). Thus, "Têll mè thĕ trúth" (the whole truth, and nothing but the truth) may be contrasted with "Têll mé thĕ trûth" (whatever you may tell other people); "bláck bîrd" (any bird black in colour) may be contrasted with "bláckbìrd" (that particular bird *Turdus merula*). The verbs "permít" and "recórd" (henceforth only primary stresses are marked) may be contrasted with their corresponding nouns "pérmit" and "récord." A feeling for antepenultimate (third syllable from the end) primary stress, revealed in such five-syllable words as equanímity, longitúdinal, notoríety, opportúnity, parsimónious, pertinácity, and vegetárian, causes stress to shift when extra syllables are added, as in "histórical," a derivative of "hístory" and "theatricálity," a derivative of "theátrical." Vowel qualities are also changed here and in such word groups as périod, periódical, periodícity; phótograph, photógraphy, photográphical. French stress may be sustained in many borrowed words; *e.g.*, bizárre, critíque, duréss, hotél, prestíge, and techníque.

Pitch, or musical tone, determined by the rate of vibration of the vocal cords, may be level, falling, rising, or falling–rising. In counting "one," "two," "three," "four," one naturally gives level pitch to each of these cardinal numerals. But if a person says "I want two, not one," he naturally gives "two" falling pitch and "one" falling–rising. In the question "One?" rising pitch is used. Word tone is called pitch, and sentence tone is referred to as intonation. The end-of-sentence cadence is important for meaning, and it therefore varies least. Three main end-of-sentence intonations can be distinguished: falling, rising, and falling–rising. Falling intonation is used in completed statements, direct commands, and sometimes in general questions unanswerable by "yes" or "no"; *e.g.*, "I have nothing to add." "Keep to the right." "Who told you that?" Rising intonation is frequently used in open-ended statements made with some reservation, in polite requests, and in particular questions answerable by "yes" or "no": "I have nothing more to say at the moment." "Let me know how you get on." "Are you sure?" The third type of end-of-sentence intonation, first falling and then rising pitch, is used in sentences that imply concessions or contrasts: "Some people do like them" (but others do not). "Don't say I didn't warn you" (because that is just what I'm now doing). Intonation is on the whole less singsong in American than in British English, and there is a narrower range of pitch. American speech may seem more monotonous but at the same time may sometimes be clearer and more readily intelligible. Everywhere English is spoken, regional dialects display distinctive patterns of intonation.

Intonation in American and British English

Morphology. *Inflection.* Modern English nouns, pronouns, and verbs are inflected. Adjectives, adverbs, prepositions, conjunctions, and interjections are invariable.

Most English nouns have plural inflection in (*-e*)*s*, but this form shows variations in pronunciation in the words cats (with a final *s* sound), dogs (with a final *z* sound), and horses (with a final *iz* sound), as also in the 3rd person singular present-tense forms of verbs: cuts (*s*), jogs (*z*), and forces (*iz*). Seven nouns have mutated (umlauted) plurals: man, men; woman, women; tooth, teeth; foot, feet; goose, geese; mouse, mice; louse, lice. Three have plurals in *-en*: ox, oxen; child, children; brother, brethren. Some remain unchanged; *e.g.*, deer, sheep, moose, grouse. Five of the seven personal pronouns have distinctive forms for subject and object.

The forms of verbs are not complex. Only the substantive verb ("to be") has eight forms: be, am, is, are, was, were, being, been. Strong verbs have five forms: ride, rides, rode, riding, ridden. Regular or weak verbs customarily have four: walk, walks, walked, walking. Some that end in a *t* or *d* have three forms only: cut, cuts, cutting. Of these three-form verbs, 16 are in frequent use.

In addition to the above inflections, English employs two other main morphological (structural) processes—affixation and composition—and two subsidiary ones—back-formation and blend.

Types of affixes

Affixation. Affixes, word elements attached to words, may either precede, as prefixes (do, undo; way, subway), or follow, as suffixes (do, doer; way, wayward). They may be native (*over*do, wayward*ness*), Greek (*hyper*bole, the*sis*), or Latin (*super*sede, pedi*ment*). Modern technologists greatly favour the neo-Hellenic prefixes *macro-* "long, large," *micro-* "small," *para-* "alongside," *poly-* "many," and the Latin *mini-*, with its antonym *maxi-*. Greek and Latin affixes have become so fully acclimatized that they can occur together in one and the same word, as, indeed, in "ac-climat-ize-d," just used, consisting of a Latin prefix plus a Greek stem plus a Greek suffix plus an English inflection. Suffixes are bound more closely than prefixes to the stems or root elements of words. Consider, for instance, the wide variety of agent suffixes in the nouns act*or*, artis*an*, dot*ard*, engin*eer*, financ*ier*, hire*ling*, magistr*ate*, merch*ant*, scient*ist*, secret*ary*, song*ster*, stud*ent*, and work*er*. Suffixes may come to be attached to stems quite fortuitously, but, once attached, they are likely to be permanent. At the same time, one suffix can perform many functions. The suffix *-er* denotes the doer of the action in the words worker, driver, and hunter; the instrument in chopper, harvester, and roller; and the dweller in Icelander, Londoner, and Trobriander. It refers to things or actions associated with the basic concept in the words breather, "pause to take breath"; diner, "dining car on a train"; and fiver, "five-pound note." In the terms disclaimer, misnomer, and rejoinder (all from French) the suffix denotes one single instance of the action expressed by the verb. Usage may prove capricious. Whereas a writer is a person, a typewriter is a machine. For some time a computer was both, but now, with the invention and extensive use of electronic apparatus, the word is no longer used of persons.

Composition. Composition, or compounding, is concerned with free forms. The primary compounds "already," "cloverleaf," and "gentleman" show the collocation of two free forms. They differ from word groups or phrases in phonology, stress, or juncture or by a combination of two or more of these. Thus, "already" differs from "all ready" in stress and juncture, "cloverleaf" from "clover leaf" in stress, and "gentleman" from "gentle man" in phonology, stress, and juncture. In describing the structure of compound words it is necessary to take into account the relation of components to each other and the relation of the whole compound to its components. These relations diverge widely in, for example, the words cloverleaf, icebreaker, breakwater, blackbird, peace-loving, and paperback. In "cloverleaf" the first component noun is attributive and modifies the second, as also in the terms aircraft, beehive, landmark, lifeline, network, and vineyard. "Icebreaker," however, is a compound made up of noun object plus agent noun, itself consisting of verb plus agent suffix, as also in the words bridgebuilder, landowner, metalworker, minelayer, and timekeeper. The next type consists of verb plus object. It is rare in English, Dutch, and German but frequent in French, Spanish, and Italian. The English "pastime" may be compared, for example, with French *passe-temps*, the Spanish *pasatiempo*, and the Italian *passatempo*. From French comes "passport," meaning "pass (*i.e.*, enter) harbour." From Italian comes "portfolio," meaning "carry leaf." Other words of this type are daredevil, scapegrace, and scarecrow. As for the "blackbird" type, consisting of attributive adjective plus noun, it occurs frequently, as in the terms bluebell, grandson, shorthand, and wildfire. The next type, composed of object noun and a present participle, as in the terms fact-finding, heart-rending (German *herzzerreissend*), life-giving (German *lebenspendend*), painstaking, and time-consuming, occurs rarely. The last type is seen in barefoot, bluebeard, hunchback, leatherneck, redbreast, and scatterbrain.

Compound words

Back-formations and blends. Back-formations and blends are becoming increasingly popular. Back-formation is the reverse of affixation, being the analogical creation of a new word from an existing word falsely assumed to be its derivative. For example, the verb "to edit" has been formed from the noun "editor" on the reverse analogy of the noun "actor" from "to act," and similarly the verbs automate, bulldoze, commute, escalate, liaise, loaf, sightsee, and televise are backformed from the nouns automation, bulldozer, commuter, escalation, liaison, loafer, sightseer, and television. From the single noun "procession" are backformed two verbs with different stresses and meanings: procéss, "to walk in procession," and prócess, "to subject food (and other material) to a special operation."

Two types of blends

Blends fall into two groups: (1) coalescences, such as "bash" from "bang" and "smash"; and (2) telescoped forms, called portmanteau words, such as "motorcade" from "motor cavalcade." In the first group are the words clash, from clack and crash, and geep, offspring of goat and sheep. To the second group belong dormobiles, or dormitory automobiles, and slurbs, or slum suburbs. A travel monologue becomes a travelogue and a telegram sent by cable a cablegram. Aviation electronics becomes avionics; biology electronics, bionics; and nuclear electronics, nucleonics. In cablese a question mark is a quark; in computerese a binary unit is a bit. In astrophysics a quasistellar source of radio energy becomes a quasar, and a pulsating star becomes a pulsar.

Simple shortenings, such as "ad" for "advertisement," have risen in status. They are listed in dictionaries side by side with their full forms. Among such fashionable abbreviations are exam, gym, lab, lib, op, spec, sub, tech, veg, and vet. Compound shortenings, after the pattern of Russian *agitprop* for *agitatsiya propaganda,* are also becoming fashionable. Initial syllables are joined as in the words Fortran, for formula (computer) translation; mascon, for massive (lunar) concentration; and Tacomsat, for Tactical Communications Satellite.

Syntax. Sentences can be classified as (1) simple, containing one clause and predication: "John knows this country"; (2) multiple or compound, containing two or more coordinate clauses: "John has been here before, and he knows this country"; and (3) complex, containing one or more main clauses and one or more subordinate clauses: "John, who has been here before, knows this country" or "Because he has been here before, John knows this country." Simple, declarative, affirmative sentences have two main patterns with five subsidiary patterns within each. Verb and complement together form the predicate. "Complement" is here used to cover both the complement and the object of traditional grammarians (see Table 23).

Table 23: Simple Sentences—First Pattern

subject	verb	complement
1. John	knows	this country
2. Science	is	organized knowledge
3. Elizabeth	becomes	queen
4. The captain	falls	sick
5. Nothing	passes	unobserved

In (1) the complement is the direct object of a transitive verb; in (2) it is a predicative nominal group forming the second component of an equation linked to the first part by the meaningless copula is; in (3) it is a predicative noun linked with the subject by the meaningful copula becomes; in (4) it is a predicative adjective; and in (5) it is a predicative past participle.

In Table 24 each sentence contains four components: subject, verb, and two complements, first and second, or inner and outer. In (6) inner and outer complements consist of indirect object (without preposition) followed by direct object; in (7) these complements are direct object and appositive noun; in (8) direct object and predicative adjective; in (9) direct object and predicative past participle; in (10) direct object and predicative infinitive.

Table 24: Simple Sentences—Second Pattern

subject	verb	inner complement	outer complement
6. John	gives	Mary	a ring
7. The sailors	make	John	captain
8. You	have kept	your record	clean
9. The driver	finds	the road	flooded
10. We	want	you	to know

One can seldom change the word order in these 10 sentences without doing something else—adding or subtracting a word, changing the meaning. There is no better way of appreciating the importance of word position than by scrutinizing the 10 frames illustrated. If, for instance, in (6) one reverses inner and outer complements, one adds "to" and says, "John gives a ring to Mary"; one does not say "John gives a ring Mary." Some verbs, such as "explain" and "say," never omit the preposition "to" before the indirect object: "John's father explained the details to his son." "He said many things to him." If, in (10), the inner and outer complements are reversed (*e.g.,* "We want to know you"), the meaning is changed as well as the structure.

Apart from these fundamental rules of word order, the principles governing the positions of adjectives, adverbs, and prepositions call for brief comment. For attributive adjectives the rule is simple: single words regularly precede the noun, and word groups follow—*e.g.,* "an unforgettable experience" but "an experience never to be forgotten." There is a growing tendency, however, to abandon this principle, to switch groups to front position, and to say "a never to be forgotten experience." In the ordering of multiple epithets, on the other hand, some new principles are seen to be slowly emerging. Attributes denoting permanent qualities stand nearest their head nouns: "long, white beard," "six-lane elevated freeway." The order in multiple attribution tends to be as follows: determiner; quantifier; adjective of quality; adjective of size, shape, or texture; adjective of colour or material; noun adjunct (if any); head noun. Examples include: "that one solid, round, oak dining table," "these many fine, large, black race horses," "those countless memorable, long, bright summer evenings."

Positions of adjectives

Adverbs are more mobile than adjectives. Nevertheless, some tentative principles seem to be at work. Adverbs of frequency tend to come immediately after the substantive verb ("You are often late"), before other verbs ("You never know"), and between auxiliaries and full verbs ("You can never tell"). In this last instance, however, American differs from British usage. Most Americans would place the adverb before the auxiliary and say "You never can tell." (In the title of his play of that name, first performed in 1899, George Bernard Shaw avowedly followed American usage.) Adverbs of time usually occur at the beginning or end of a sentence, seldom in the middle. Particular expressions normally precede more general ones: "Neil Armstrong set foot on the Moon at 4 o'clock in the morning on July 21, 1969." An adverb of place or direction follows a verb with which it is semantically bound: "We arrived home after dark." Other adverbs normally take end positions in the order of manner, place, and time: "Senator Smith summed it all up most adroitly [manner] in Congress [place] last night [time]."

In spite of its etymology (Latin *prae-positio* "before placing"), a preposition may sometimes follow the noun it governs, as in "all the world over," "the clock round," and "the whole place through." "This seems a good place to live in" seems more natural to most speakers than "This seems a good place in which to live." "Have you anything to open this can with?" is now more common than "Have you anything with which to open this can?"

The above are principles rather than rules, and in the end it must be agreed that English syntax lacks regimentation. Its structural laxity makes English an easy language to speak badly. It also makes English prone to ambiguity. "When walking snipe always approach up wind," a shooting manual directs. The writer intends the reader to understand, "When you are walking to flush snipe always approach them up against the wind." "John kept the car in the garage" can mean either (1) "John retained that car you see in the garage, and sold his other one" or (2) "John housed the car in the garage, and not elsewhere." "Flying planes can be dangerous" is ambiguous because it may mean either (1) "Planes that fly can be dangerous" or (2) "It is dangerous to fly planes."

Two ways in which "John gives Mary a ring" can be stated in the passive are: (1) "A ring is given to Mary by John" and (2) "Mary is given a ring by John." Concerning this same action, four types of question can be formulated: (1) "Who gives Mary a ring?" The information sought is the identity of the giver. (2) "Does John give Mary a ring?" The question may be answered by "yes" or "no." (3) "John gives Mary a ring, doesn't he?" Confirmation is sought of the questioner's belief that John does in fact give Mary a ring. (4) "John gives Mary a ring?" This form, differing from the declarative statement only by the question mark in writing, or by rising intonation in speech, calls, like sentences (2) and (3), for a "yes" or "no" answer but suggests doubt on the part of the questioner that the action is taking place.

Origins of English words

Vocabulary. The vocabulary of Modern English is approximately half Germanic (Old English and Scandinavian) and half Italic or Romance (French and Latin), with copious and increasing importations from Greek in science and technology and with considerable borrowings from Dutch, Low German, Italian, Spanish, German, Arabic, and many other languages. Names of basic concepts and things come from Old English or Anglo-Saxon: heaven and earth, love and hate, life and death, beginning and end, day and night, month and year, heat and cold, way and path, meadow and stream. Cardinal numerals come from Old English, as do all the ordinal numerals except "second" (Old English *other,* which still retains its older meaning in "every other day"). "Second" comes from Latin *secundus* "following," through French *second,* related to Latin *sequi* "to follow," as in English "sequence." From Old English come all the personal pronouns (except "they," "their," and "them," which are from Scandinavian), the auxiliary verbs (except the marginal "used," which is from French), most simple prepositions, and all conjunctions.

Numerous nouns would be identical whether they came from Old English or Scandinavian: father, mother, brother (but not sister); man, wife; ground, land, tree, grass; summer, winter; cliff, dale. Many verbs would also be identical, especially monosyllabic verbs—bring, come, get, hear, meet, see, set, sit, spin, stand, think. The same is true of the adjectives full and wise; the colour names gray, green, and white; the disjunctive possessives mine and thine (but not ours and yours); the terms north and west (but not south and east); and the prepositions over and under. Just a few English and Scandinavian doublets coexist in current speech: no and nay, yea and ay, from and fro, rear (*i.e.,* to bring up) and raise, shirt and skirt (both related to the adjective short), less and loose. From Scandinavian, "law" was borrowed early, whence "bylaw," meaning "village law," and "outlaw," meaning "man outside the law." "Husband" (*hus-bondi*) meant "householder," whether single or married, whereas "fellow" (*fe-lagi*) meant one who "lays fee" or shares property with another, and so "partner, shareholder." From Scandinavian come the common nouns axle (tree), band, birth, bloom, crook, dirt, egg, gait, gap, girth, knife, loan, race, rift, root, score, seat, skill, sky, snare, thrift, and window; the adjectives awkward, flat, happy, ill, loose, rotten, rugged, sly, tight, ugly, weak, and wrong; and many verbs, including call, cast, clasp, clip, crave, die, droop, drown, flit, gape, gasp, glitter, life, rake, rid, scare, scowl, skulk, snub, sprint, thrive, thrust, and want.

Borrowing from French

The debt of the English language to French is large. The terms president, representative, legislature, congress, constitution, and parliament are all French. So, too, are duke, marquis, viscount, and baron; but king, queen, lord, lady, earl, and knight are English. City, village, court, palace, manor, mansion, residence, and domicile are French; but town, borough, hall, house, bower, room, and home are English. Comparison between English and French synonyms shows that the former are more human and concrete, the latter more intellectual and abstract; *e.g.,* the terms freedom and liberty, friendship and amity, hatred and enmity, love and affection, likelihood and probability, truth and veracity, lying and mendacity. The superiority of French cooking is duly recognized by the adoption of such culinary terms as boil, broil, fry, grill, roast, souse, and toast. "Breakfast" is English, but "dinner" and "supper" are French. "Hunt" is English, but "chase," "quarry," "scent," and "track" are French. Craftsmen bear names of English origin: baker, builder, fisher (man), hedger, miller, shepherd, shoemaker, wainwright, and weaver, or webber. Names of skilled artisans, however, are French: carpenter, draper, haberdasher, joiner, mason, painter, plumber, and tailor. Many terms relating to dress and fashion, cuisine and viniculture, politics and diplomacy, drama and literature, art and ballet come from French.

Greek borrowings

In the spheres of science and technology many terms come from Classical Greek through French, or directly from Greek. Pioneers in research and development now regard Greek as a kind of inexhaustible quarry from which they can draw linguistic material at will. By prefixing the Greek adverb *tēle* "far away, distant" to the existing compound photography, "light writing," they create the precise term "telephotography" to denote the photographing of distant objects by means of a special lens. By inserting the prefix *micro-* "small" into this same compound, they make the new term "photomicrography," denoting the electronic photographing of bacteria and viruses. Such neo-Hellenic derivatives would probably have been unintelligible to Plato and Aristotle. Many Greek compounds and derivatives have Latin equivalents with slight or considerable differentiations in meaning (see Table 25).

Table 25: Equivalent Compounds and Derivatives*

from the Greek	from the Latin
Nouns	
dys-troph-y	mal-nutr-it-ion
hypo-sta-sis	sub-stan-ce
hypo-the-sis	sup-pos-it-ion
meta-morph-o-sis	trans-form-at-ion
meta-phor	trans-fer
meta-the-sis	trans-pos-it-ion
peri-pher-y	circum-fer-en-ce
peri-phra-sis	circum-loc-ut-ion
sym-path-y	com-pass-ion
syn-drom-e	con-curr-en-ce
syn-op-sis	con-spect-us
syn-the-sis	com-pos-it-ion
sy-zyg-y	con-junc-t-ion
Adjectives	
dia-phan-*ous*	trans-par-ent
hyper-aesth-et-ic	super-sens-it-ive
hyper-phys-ic-*al*	super-nat-ur-al
hypo-derm-ic	sub-cut-an-eous
hypo-ge-*al*	sub-terr-an-ean
melan-chol-ic	atra-bil-ious
mono-morph-ic	uni-form
oxy-phyll-*ous*	acut-i-fol-i-ate
peri-pat-et-ic	circum-amb-ul-at-ory
phos-phor-*escent*	lumin-i-fer-ous
poly-glott-*al*	multi-lingu-al
sphen-oid	cunei-form
syn-chron-ic	con-temp-or-ary

*The italicized suffixes *-al*, *-escent*, and *-ous*, attached to some of the Greek adjectives, are of Latin origin.

At first sight it might appear that some of these equivalents, such as "metamorphosis" and "transformation," are sufficiently synonymous to make one or the other redundant. In fact, however, "metamorphosis" is more technical and therefore more restricted than "transformation." In mythology it signifies a magical shape changing; in nature it denotes a postembryonic development such as that of a tadpole into a frog, a cocoon into a silkworm, or a chrysalis into a butterfly. Transformation, on the other hand, means any kind of change from one state to another.

Ever since the 12th century, when merchants from the Netherlands made homes in East Anglia, Dutch words have infiltrated into Midland speech. For centuries a form of Low German was used by seafaring men in North Sea ports. Old nautical terms still in use include buoy, deck, dock, freebooter, hoist, leak, pump, skipper, and yacht. The Dutch in New Amsterdam (later New York) and adjacent settlements gave the words boss, cookie, dope, snoop, and waffle to American speech. The Dutch in Cape Province gave the terms apartheid, commandeer, commando, spoor, and trek to South African speech.

German contributions

The contribution of High German has been on a different level. In the 18th and 19th centuries it lay in technicalities of geology and mineralogy and in abstractions relating to literature, philosophy, and psychology. In the 20th century this contribution has sometimes been indirect. "Unclear" and "meaningful" echoed German *unklar* and *bedeutungsvoll,* or *sinnvoll.* "Ring road" (a British term applied to roads encircling cities or parts of cities) translated *Ringstrasse;* "round trip," *Rundfahrt;* and "the turn of the century," *die Jahrhundertwende.* The terms "classless society," "inferiority complex," and "wishful thinking" echoed *die klassenlöse Gesellschaft, der Minderwertigkeitskomplex,* and *das Wunschdenken.*

Along with the rest of the Western world, English has accepted Italian as the language of music. The names of voices, parts, performers, instruments, forms of composition, and technical directions are all Italian. Many of the latter—allegro, andante, cantabile, crescendo, diminuendo, legato, maestoso, obbligato, pizzicato, staccato, and vibrato—are also used metaphorically. In architecture, the terms belvedere, corridor, cupola, grotto, pedestal, pergola, piazza, pilaster, and rotunda are accepted; in literature, burlesque, canto, extravaganza, stanza, and many more are used.

From Spanish, English has acquired the words armada, cannibal, cigar, galleon, guerrilla, matador, mosquito, quadroon, tornado, and vanilla, some of these loanwords going back to the 16th century, when sea dogs encountered hidalgos on the high seas. Many names of animals and plants have entered English from indigenous languages through Spanish: "potato" through Spanish *patata* from Taino *batata,* and "tomato" through Spanish *tomate* from Nahuatl *tomatl.* Other words have entered from Latin America by way of Texas, New Mexico, Arizona, and California; *e.g.,* such words as canyon, cigar, estancia, lasso, mustang, pueblo, and rodeo. Some have gathered new connotations: bonanza, originally denoting "goodness," came through miners' slang to mean "spectacular windfall, prosperity"; mañana, "tomorrow," acquired an undertone of mysterious unpredictability.

From Arabic through European Spanish, through French from Spanish, through Latin, or occasionally through Greek, English has obtained the terms alchemy, alcohol, alembic, algebra, alkali, almanac, arsenal, assassin, attar, azimuth, cipher, elixir, mosque, nadir, naphtha, sugar, syrup, zenith, and zero. From Egyptian Arabic, English has recently borrowed the term loofah (also spelled luffa). From Hebrew, directly or by way of Vulgate Latin, come the terms amen, cherub, hallelujah, manna, messiah, pharisee, rabbi, sabbath, and seraph; jubilee, leviathan, and shibboleth; and, more recently, kosher, and kibbutz.

English has freely adopted and adapted words from many other languages, acquiring them sometimes directly and sometimes by devious routes. Each word has its own history. The following lists indicate the origins of a number of English words: Welsh—flannel, coracle, cromlech, eisteddfod; Cornish—gull, brill, dolmen; Gaelic and Irish—shamrock, brogue, leprechaun, ogham, Tory, galore, blarney, hooligan, clan, claymore, bog, plaid, slogan, sporran, cairn, whisky, pibroch; Breton—menhir, penguin; Norwegian—ski, ombudsman; Finnish—sauna; Russian—kvass, ruble, tsar, verst, mammoth, ukase, astrakhan, vodka, samovar, tundra (from Lapp), troika, pogrom, duma, soviet, bolshevik, intelligentsia (from Latin through Polish), borscht, balalaika, sputnik, soyuz, salyut, lunokhod; Polish—mazurka; Czech—robot; Hungarian—goulash, paprika; Portuguese—marmalade, flamingo, molasses, veranda, port (wine), dodo; Basque—bizarre; Turkish—janissary, turban, coffee, kiosk, caviar, pasha, odalisque, fez, bosh; Hindi—nabob, guru, sahib, maharajah, mahatma, pundit, punch (drink), juggernaut, cushy, jungle, thug, cheetah, shampoo, chit, dungaree, pucka, gymkhana, mantra, loot, pajamas, dinghy, polo; Persian—paradise, divan, purdah, lilac, bazaar, shah, caravan, chess, salamander, taffeta, shawl, khaki; Tamil—pariah, curry, catamaran, mulligatawny; Chinese—tea (Amoy), sampan; Japanese—shogun, kimono, mikado, tycoon, hara-kiri, gobang, judo, jujitsu, bushido, samurai, banzai, tsunami, satsuma, No (the dance drama), karate, Kabuki; Malay—ketchup, sago, bamboo, junk, amuck, orangutan, compound (fenced area), raffia; Polynesian—taboo, tattoo; Hawaiian—ukulele; African languages—chimpanzee, goober, mumbo jumbo, voodoo; Eskimo—kayak, igloo, anorak, mukluk; Algonkian—totem; Nahuatl—mescal; languages of the Caribbean—hammock, hurricane, tobacco, maize, iguana; Aboriginal Australian—kangaroo, corroboree, wallaby, wombat, boomerang, paramatta, budgerigar.

The alphabet

Orthography. The Latin alphabet originally had 20 letters, the present English alphabet minus *J, K, V, W, Y,* and *Z.* The Romans themselves added *K* for use in abbreviations and *Y* and *Z* in words transcribed from Greek. After its adoption by the English, this 23-letter alphabet developed *W* as a ligatured doubling of *U* and later *J* and *V* as consonantal variants of *I* and *U.* The resultant alphabet of 26 letters has both uppercase, or capital, and lowercase, or small, letters.

English spelling is based for the most part on that of the 15th century, but pronunciation has changed considerably since then, especially that of long vowels and diphthongs. The extensive change in the pronunciation of vowels, known as the Great Vowel Shift, affected all of Geoffrey Chaucer's seven long vowels, and for centuries spelling remained untidy. If the meaning of the message was clear, the spelling of individual words seemed unimportant. In the 17th century during the English Civil War, compositors adopted fixed spellings for practical reasons, and in the order-loving 18th century uniformity became more and more fashionable. Since Samuel Johnson's *Dictionary of the English Language* (1755), orthography has remained fairly stable. Numerous tacit changes, such as "music" for "musick" (*c.* 1880) and "fantasy" for "phantasy" (*c.* 1920), have been accepted, but spelling has nevertheless continued to be in part unphonetic. Attempts have been made at reform. Indeed, every century has had its reformers since the 13th, when an Augustinian canon named Orm devised his own method of differentiating short vowels from long by doubling the succeeding consonants or, when this was not feasible, by marking short vowels with a superimposed breve mark (˘). William Caxton, who set up his wooden printing press at Westminster in 1476, was much concerned with spelling problems throughout his working life. Noah Webster produced his *Spelling Book,* in 1783, as a precursor to the first edition (1828) of his *American Dictionary of the English Language.* The 20th century has produced many zealous reformers. Three systems, supplementary to traditional spelling, are actually in use for different purposes: (1) the Initial Teaching (Augmented Roman) Alphabet (ITA) of 44 letters used by educationists in the teaching of children under seven; (2) the Shaw alphabet of 48 letters, designed in implementation of the will of George Bernard Shaw; and (3) the International Phonetic Alphabet (IPA), constructed on the basis of one symbol for one individual sound and used by many trained linguists. Countless other systems have been worked out from time to time, of which R.E. Zachrisson's "Anglic" (1930) and Axel Wijk's *Regularized English* (1959) may be the best.

Obstacles to spelling reform

Meanwhile, the great publishing houses continue unperturbed because drastic reform remains impracticable, undesirable, and unlikely. This is because there is no longer one criterion of correct pronunciation but several standards throughout the world; regional standards are themselves not static, but changing with each new generation; and, if spelling were changed drastically, all the books in English in the world's public and private libraries would become inaccessible to readers without special study.

HISTORICAL BACKGROUND

Among highlights in the history of the English language, the following stand out most clearly: the settlement in Britain of Jutes, Saxons, and Angles in the 5th and 6th centuries; the arrival of St. Augustine in 597 and the subsequent conversion of England to Latin Christianity; the Viking invasions of the 9th century; the Norman Conquest of 1066; the Statute of Pleading in 1362 (this required that court proceedings be conducted in English); the setting up of Caxton's printing press at Westminster in 1476; the full flowering of the Renaissance in the 16th century; the publishing of the King James Bible in 1611; the completion of Johnson's *Dictionary* of 1755; and the expansion to North America and South Africa in the 17th century and to India, Australia, and New Zealand in the 18th.

Old English. The Jutes, Angles, and Saxons lived in Jutland, Schleswig, and Holstein, respectively, before settling in Britain. According to the Venerable Bede, the first historian of the English people, the first Jutes, Hengist and Horsa, landed at Ebbsfleet in the Isle of Thanet in 449; and the Jutes later settled in Kent, southern Hampshire, and the Isle of Wight. The Saxons occupied the rest of England south of the Thames, as well as modern Middle-sex and Essex. The Angles eventually took the remainder of England as far north as the Firth of Forth, including the future Edinburgh and the Scottish Lowlands. In both Latin and Common Germanic the Angles' name was Angli, later mutated in Old English to Engle (nominative) and Engla (genitive). "Engla land" designated the home of all three tribes collectively, and both King Alfred (known as Alfred the Great) and Abbot Aelfric, author and grammarian, subsequently referred to their speech as Englisc. Nevertheless, all the evidence indicates that Jutes, Angles, and Saxons retained their distinctive dialects.

The River Humber was an important boundary, and the Anglian-speaking region developed two speech groups: to the north of the river, Northumbrian, and, to the south, Southumbrian, or Mercian. There were thus four dialects: Northumbrian, Mercian, West Saxon, and Kentish (see Figure 13). In the 8th century, Northumbrian led in lit-

Four Old English dialects

Figure 13: Old English dialects.

erature and culture, but that leadership was destroyed by the Viking invaders, who sacked Lindisfarne, an island near the Northumbrian mainland, in 793. They landed in strength in 865. The first raiders were Danes, but they were later joined by Norwegians from Ireland and the Western Isles who settled in modern Cumberland, Westmorland, northwest Yorkshire, Lancashire, north Cheshire, and the Isle of Man. In the 9th century, as a result of the Norwegian invasions, cultural leadership passed from Northumbria to Wessex. During King Alfred's reign, in the last three decades of the 9th century, Winchester became the chief centre of learning. There the Parker Chronicle (a manuscript of the Anglo-Saxon Chronicle) was written; there the Latin works of the priest and historian Paulus Orosius, St. Augustine, St. Gregory, and the Venerable Bede were translated; and there the native poetry of Northumbria and Mercia was transcribed into the West Saxon dialect. This resulted in West Saxon's becoming "standard Old English"; and later, when Aelfric (*c.* 955–*c.* 1010) wrote his lucid and mature prose at Winchester, Cerne Abbas, and Eynsham, the hegemony of Wessex was strengthened.

In standard Old English, adjectives were inflected as well as nouns, pronouns, and verbs. Nouns were inflected for four cases (nominative, genitive, dative, and accusative) in singular and plural. Five nouns of first kinship—*faeder, mōdor, brōthor, sweostor,* and *dohtor* ("father," "mother," "brother," "sister," and "daughter," respectively)—had their own set of inflections. There were 25 nouns such as *mon, men* ("man," "men") with mutated, or umlauted, stems. Adjectives had strong and weak declensions, the strong showing a mixture of noun and pronoun endings and the weak following the pattern of weak nouns. Personal, possessive, demonstrative, interrogative, indefinite, and relative pronouns had full inflections. The pronouns of the 1st and 2nd persons still had distinctive dual forms:

iċ	"I"	*wit*	"we two"	*wē*	"we"
thū (*þū*)	"thou"	*ġit*	"you two"	*ġē*	"you"

There were two demonstratives: *sē, sēo, thaet,* meaning "that," and *thes, thēos, this,* meaning "this," but no articles, the definite article being expressed by use of the demonstrative for "that" or not expressed at all. Thus, "the good man" was *sē gōda mon* or plain *gōd mon.* The function of the indefinite article was performed by the numeral *ān* "one" in *ān mon* "a man," by the adjective–pronoun *sum* in *sum mon* "a (certain) man," or not expressed, as in *thū eart gōd mon* "you are a good man."

Verbs had two tenses only (present–future and past), three moods (indicative, subjunctive, and imperative), two numbers (singular and plural), and three persons (1st, 2nd, and 3rd). There were two classes of verb stems. (A verb stem is that part of a verb to which inflectional changes—changes indicating tense, mood, number, etc.—are added.) One type of verb stem, called vocalic because an internal vowel shows variations, is exemplified by the verb for "sing": *singan, singth, sang, sungon, gesungen.* The word for "deem" is an example of the other, called consonantal: *dēman, dēmth, dēmde, dēmdon, gedēmed.* Such verbs are called strong and weak, respectively.

Inflection of verbs

All new verbs, whether derived from existing verbs or from nouns, belonged to the consonantal type. Some verbs of great frequency (antecedents of the modern words "be," "shall," "will," "do," "go," "can," "may," and so on) had their own peculiar patterns of inflections.

Grammatical gender persisted throughout the Old English period. Just as Germans now say *der Fuss, die Hand,* and *das Auge* (masculine, feminine, and neuter terms for "the foot," "the hand," and "the eye"), so, for these same structures, Aelfric said *sē fōt, sēo hond,* and *thaet ēaġe,* also masculine, feminine, and neuter. The three words for "woman," *wīfmon, cwene,* and *wīf,* were masculine, feminine, and neuter, respectively. *Hors* "horse," *sċēap* "sheep," and *maeġden* "maiden" were all neuter. *Eorthe* "earth" was feminine, but *lond* "land" was neuter. *Sunne* "sun" was feminine, but *mōna* "moon" was masculine. This simplification of grammatical gender resulted from the fact that the gender of Old English substantives was not always indicated by the ending but rather by the terminations of the adjectives and demonstrative pronouns used with the substantives. When these endings were lost, all outward marks of gender disappeared with them. Thus, the weakening of inflections and loss of gender occurred together. In the North, where inflections weakened earlier, the marks of gender likewise disappeared first. They survived in the South as late as the 14th century.

Because of the greater use of inflections in Old English, word order was freer than today. The sequence of subject, verb, and complement was normal, but when there were outer and inner complements the second was put in the dative case after *to: Sē biscop hālgode Ēadrēd tō cyninge* "The bishop consecreated Edred king." After an introductory adverb or adverbial phrase the verb generally took second place as in modern German: *Nū bydde iċ ān thing* "Now I ask [literally, "ask I"] one thing"; *Thȳ ilcan ġēare gesette Aelfrēd cyning Lundenburg* "In that same year Alfred the king occupied London." Impersonal verbs had no subject expressed. Infinitives constructed with auxiliary verbs were placed at the ends of clauses or sentences: *Hīe ne dorston forth bī th ǣre ēa siglan* "They dared not sail beyond that river" (*siglan* is the infinitive); *Iċ wolde thās lytlan bōc āwendan* "I wanted to translate this little book" (*āwendan* is the infinitive). The verb usually came last in a dependent clause—*e.g., āwrītan wile* in *gif hwā thās bōc āwrītan wile* (*gerihte hē hīe be th ǣre bysene*) "If anyone wants to copy this book (let him correct his copy by the original)." Prepositions (or postpositions) frequently followed their objects. Negation was often repeated for emphasis.

Figure 14: Middle English dialects.

Middle English. One result of the Norman Conquest of 1066 was to place all four Old English dialects more or less on a level. West Saxon lost its supremacy and the centre of culture and learning gradually shifted from Winchester to London. The old Northumbrian dialect became divided into Scottish and Northern, although little is known of either of these divisions before the end of the 13th century (Figure 14). The old Mercian dialect was split into East and West Midland. West Saxon became slightly diminished in area and was more appropriately named the South Western dialect. The Kentish dialect was considerably extended and was called South Eastern accordingly. All five Middle English dialects (Northern, West Midland, East Midland, South Western, and South Eastern) went their own ways and developed their own characteristics. The so-called Katherine Group of writings (1180–1210), associated with Hereford, a town not far from the Welsh border, adhered most closely to native traditions, and there is something to be said for regarding this West Midland dialect, least disturbed by French and Scandinavian intrusions, as a kind of Standard English in the High Middle Ages.

Norman changes in spelling

Another outcome of the Norman Conquest was to change the writing of English from the clear and easily readable insular hand of Irish origin to the delicate Carolingian script then in use on the Continent. With the change in appearance came a change in spelling. Norman scribes wrote Old English *y* as *u*, *ȳ* as *ui*, *ū* as *ou* (*ow* when final). Thus, *mycel* ("much") appeared as *muchel*, *fȳr* ("fire") as *fuir*, *hūs* ("house") as *hous*, and *hū* ("how") as *how*. For the sake of clarity (*i.e.*, legibility) *u* was often written *o* before and after *m*, *n*, *u*, *v*, and *w*; and *i* was sometimes written *y* before and after *m* and *n*. So *sunu* ("son") appeared as *sone* and *him* ("him") as *hym*. Old English *cw* was changed to *qu*; *hw* to *wh*, *qu*, or *quh*; *ċ* to *ch* or *tch*; *sċ* to *sh*; *-ċġ-* to *-gg-*; and *-ht* to *ght*. So Old English *cwēn* appeared as *queen*; *hwaet* as *what*, *quat*, or *quhat*; *dīċ* as *ditch*; *sċip* as *ship*; *secge* as *segge*; and *miht* as *might*.

For the first century after the Conquest, most loanwords came from Normandy and Picardy, but with the extension south to the Pyrenees of the Angevin empire of Henry II (reigned 1154–89), other dialects, especially Central French, or Francien, contributed to the speech of the aristocracy. As a result, Modern English acquired the forms canal, catch, leal, real, reward, wage, warden, and warrant from Norman French side by side with the corresponding forms channel, chase, loyal, royal, regard, gage, guardian, and guarantee, from Francien. King John lost Normandy in 1204. With the increasing power of the Capetian kings of Paris, Francien gradually predominated. Meanwhile, Latin stood intact as the language of learning. For three centuries, therefore, the literature of England was trilingual. *Ancrene Riwle*, for instance, a guide or rule (*riwle*) of rare quality for recluses or anchorites (*ancren*), was disseminated in all three languages.

Final mute *e*

The sounds of the native speech changed slowly. Even in late Old English short vowels had been lengthened before *ld*, *rd*, *mb*, and *nd*, and long vowels had been shortened before all other consonant groups and before double consonants. In early Middle English short vowels of whatever origin were lengthened in the open stressed syllables of disyllabic words. An open syllable is one ending in a vowel. Both syllables in Old English *nama* "name," *mete* "meat, food," *nosu* "nose," *wicu* "week," and *duru* "door" were short, and the first syllables, being stressed, were lengthened to *nāme*, *mēte*, *nōse*, *wēke*, and *dōre* in the 13th and 14th centuries. A similar change occurred in 4th-century Latin, in 13th-century German, and at different times in other languages. The popular notion has arisen that final mute *-e* in English makes a preceding vowel long; in fact, it is the lengthening of the vowel that has caused *e* to be lost in pronunciation. On the other hand, Old English long vowels were shortened in the first syllables of trisyllabic words, even when those syllables were open; *e.g.*, *hāligdaeg* "holy day," *ǣrende* "message, errand," *crīstendōm* "Christianity," and *sūtherne* "southern," became *hŏliday* (Northern *hăliday*), *ĕrrende*, *chrĭstendom*, and *sŭtherne*. This principle still operates in current English. Compare, for example, trisyllabic derivatives such as the words chastity, criminal, fabulous, gradual, gravity, linear, national, ominous, sanity, and tabulate with the simple nouns and adjectives chaste, crime, fable, grade, grave, line, nation, omen, sane, and table.

There were significant variations in verb inflections in the Northern, Midland, and Southern dialects as shown in Table 26. The Northern infinitive was already one syllable (*sing* rather than the Old English *singan*), whereas the past participle *-en* inflection of Old English was strictly kept. These apparently contradictory features can be attributed entirely to Scandinavian, in which the final *-n* of the infinitive was lost early in *singa*, and the final *-n* of the past participle was doubled in *sunginn*. The Northern unmutated present participle in *-and* was also of Scandinavian origin. Old English mutated *-ende* (German *-end*) in the present participle had already become *-inde* in late West Saxon (Southern in Table 26), and it was this Southern *-inde* that blended with the *-ing* suffix (German *-ung*) of nouns of action that had already become near-gerunds in such compound nouns as *athswering* "oath swearing" and *writingfether* "writing feather, pen." This blending of present participle and gerund was further helped by the fact that Anglo-Norman and French *-ant* was itself a coalescence of Latin present participles in *-antem*, *-entem*, and Latin gerunds in *-andum*, *-endum*. The Northern second person singular *singis* was inherited unchanged from Common Germanic. The final *t* sound in Midland *-est* and Southern *-st* was excrescent, comparable with the final *t* in modern "amidst" and "amongst" from older *amiddes* and *amonges*. The Northern third person singular *singis* had a quite different origin. Like the *singis* of the plural, it resulted almost casually from an inadvertent retraction of the tongue in enunciation from an interdental *-th* sound to postdental *-s*. Today the form "singeth" survives as a

Table 26: Variations in Verb Inflections

	Northern	Midland	Southern
Infinitive	sing	singe(n)	singen
Present participle	singand	singende	singinde
Present singular			
1st person	singe	singe	singe
2nd person	singis	singes(t)	singst
3rd person	singis	singeth-es	singeth
Present plural	singis	singen	singeth
Past participle	sungen	(y)sunge(n)	ysunge

poetic archaism. Shakespeare used both *-eth* and *-s* endings ("It [mercy] blesseth him that gives and him that takes," *The Merchant of Venice*). The Midland present plural inflection *-en* was taken from the subjunctive. The past participle prefix *y-* developed from the Old English perfective prefix *ge-*.

Chaucer's language

Chaucer, who was born and died in London, spoke a dialect that was basically East Midland. Compared with his contemporaries, he was remarkably modern in his use of language. He was in his early 20s when the Statute of Pleading (1362) was passed, by the terms of which all court proceedings were henceforth to be conducted in English, though "enrolled in Latin." Chaucer himself used four languages; he read Latin (Classical and Medieval) and spoke French and Italian on his travels. For his own literary work he deliberately chose English.

Transition from Middle English to Early Modern English. The death of Chaucer at the close of the century (1400) marked the beginning of the period of transition from Middle English to the Early Modern English stage. The Early Modern English period is regarded by many scholars as beginning in about 1500 and terminating with the return of the monarchy (John Dryden's *Astraea Redux*) in 1660. The 15th century witnessed three outstanding developments: the rise of London English, the invention of printing, and the spread of the new learning.

The speech of London

Although the population of London in 1400 was only about 40,000, it was by far the largest city in England. York came second, followed by Bristol, Coventry, Plymouth, and Norwich. The Midlands and East Anglia, the most densely peopled parts of England, supplied London with streams of young immigrants. The speech of the capital was mixed, and it was changing. The seven long vowels of Chaucer's speech had already begun to shift. Incipient diphthongization of high front /i:/ (the *ee* sound in "meet") and high back /u:/ (as in "fool") led to instability in the other five long vowels. (Symbols within slash marks are taken from the International Phonetic Alphabet.) This remarkable event, known as the Great Vowel Shift, changed the whole vowel system of London English. As /i:/ and /u:/ became diphthongized to /ai/ (as in "bide") and /au/ (as in "house") respectively, so the next highest vowels, /e:/ (this sound can be heard in the first part of the diphthong in "name") and /o:/ (a sound that can be heard in the first part of the diphthong in "home"), moved up to take their places, and so on. The whole process is summarized in Table 27. Every one of the sounds appearing in this table can still be heard somewhere in living English dialects.

Table 27: Vowel Shifts in London English

Chaucer's spelling	Chaucer's pronunciation*	Shakespeare's pronunciation	present pronunciation*	present spelling
lyf	li:f	leif	laif	life
deed	de:d	di:d	di:d	deed
deel	dɛ:l	de:l	di:l	deal
name	na:mə	nɛ:m	neim	name
hoom	hɔ:m	ho:m	houm	home
mone	mo:n	mu:n	mu:n	moon
hous	hu:s	hous	haus	house

*Expressed in the International Phonetic Alphabet.

When Caxton started printing at Westminster in the late summer of 1476, he was painfully aware of the uncertain state of the English language. In his prologues and epilogues to his translations he made some revealing observations on the problems that he had encountered as translator and editor. At this time, sentence structures were being gradually modified, but many remained untidy. For the first time, nonprofessional scribes, including women, were writing at length.

The revival of classical learning was one aspect of that Renaissance, or spiritual rebirth, that arose in Italy and spread to France and England. It evoked a new interest in Greek on the part of learned men such as William Grocyn and Thomas Linacre, Sir Thomas More and Desiderius Erasmus. John Colet, dean of St. Paul's in the first quarter of the 16th century, startled his congregation by expounding the Pauline Epistles of the New Testament as living letters. The deans who had preceded him had known no Greek, because they had found in Latin all that they required. Only a few medieval churchmen, such as Robert Grosseteste, bishop of Lincoln, and the Franciscan Roger Bacon could read Greek with ease. The names of the seven liberal arts of the medieval curricula (the trivium and the quadrivium), it is true, were all Greek—grammar, logic, and rhetoric; arithmetic, geometry, astronomy, and music—but they had come into English by way of French.

Renaissance scholars adopted a liberal attitude to language. They borrowed Latin words through French, or Latin words direct; Greek words through Latin, or Greek words direct. Latin was no longer limited to Church Latin: it embraced all Classical Latin. For a time the whole Latin lexicon became potentially English. Some words, such as consolation and infidel, could have come from either French or Latin. Others, such as the terms abacus, arbitrator, explicit, finis, gratis, imprimis, item, memento, memorandum, neuter, simile, and videlicet, were taken straight from Latin. Words that had already entered the language through French were now borrowed again, so that doublets arose: benison and benediction; blame and blaspheme; chance and cadence; count and compute; dainty and dignity; frail and fragile; poor and pauper; purvey and provide; ray and radius; sever and separate; strait and strict; sure and secure. The Latin adjectives for "kingly" and "lawful" have even given rise to triplets; in the forms real, royal, and regal and leal, loyal, and legal, they were imported first from Anglo-Norman, then from Old French, and last from Latin direct.

After the dawn of the 16th century, English prose moved swiftly toward modernity. In 1525 Lord Berners completed his translation of Jean Froissart's *Chronicle*, and William Tyndale translated the New Testament. One-third of the King James Bible (1611), it has been computed, is worded exactly as Tyndale left it; and between 1525 and 1611 lay the Tudor Golden Age, with its culmination in Shakespeare. Too many writers, to be sure, used "inkhorn terms," newly-coined, ephemeral words, and too many vacillated between Latin and English. Sir Thomas More actually wrote his *Utopia* in Latin. It was translated into French during his lifetime but not into English until 1551, some years after his death. Francis Bacon published *De dignitate et augmentis scientiarum (On the Dignity and Advancement of Learning*, an expansion of his earlier *Advancement of Learning)* in Latin in 1623. William Harvey announced his epoch-making discovery of the circulation of the blood in his Latin *De Motu Cordis et Sanguinis in Animalibus* (1628; *On the Motion of the Heart and Blood in Animals*). John Milton composed polemical treatises in the language of Cicero. As Oliver Cromwell's secretary, he corresponded in Latin with foreign states. His younger contemporary Sir Isaac Newton lived long enough to bridge the gap. He wrote his *Principia* (1687) in Latin but his *Opticks* (1704) in English.

Plans to regulate English

Restoration period. With the restoration of the monarchy in 1660, men again looked to France. John Dryden admired the Académie Française and greatly deplored that the English had "not so much as a tolerable dictionary, or a grammar; so that our language is in a manner barbarous" as compared with elegant French. After the passionate controversies of the Civil War, this was an age of cool scientific nationalism. In 1662 the Royal Society of London for the Promotion of Natural Knowledge received its charter. Its first members, much concerned with language, appointed a committee of 22 "to improve the English tongue particularly for philosophic purposes." It included Dryden, the diarist John Evelyn, Bishop Thomas Sprat, and the poet Edmund Waller. Sprat pleaded for "a close, naked, natural way of speaking; positive expressions; clear senses, a native easiness; bringing all things as near the mathematical plainness" as possible. The committee, however, achieved no tangible result, and failed in its attempt to found an authoritative arbiter over the English tongue. A second attempt was made in 1712, when Jonathan Swift addressed an open letter to Robert Harley, earl of Oxford, then Lord Treasurer, making "A Proposal for Correcting, Improving, and Ascertaining [fixing] the English Tongue."

This letter received some popular support, but its aims were frustrated by a turn in political fortunes. Queen Anne died in 1714. The Earl of Oxford and his fellow Tories, including Swift, lost power. No organized attempt to found a language academy on French lines has ever been made since.

With Dryden and Swift the English language reached its full maturity. Their failure to found an academy was partly counterbalanced by Samuel Johnson in his *Dictionary* (published in 1755) and by Robert Lowth in his *Grammar* (published in 1761).

Johnson's *Dictionary*

Age of Johnson. In the making of his *Dictionary*, Johnson took the best conversation of contemporary London and the normal usage of reputable writers after Sir Philip Sidney (1554–86) as his criteria. He exemplified the meanings of words by illustrative quotations. Johnson admitted that "he had flattered himself for a while" with "the prospect of fixing our language" but that thereby "he had indulged expectation which neither reason nor experience could justify." The two-folio work of 1755 was followed in 1756 by a shortened, one-volume version that was widely used far into the 20th century. Revised and enlarged editions of the unabbreviated version were made by Archdeacon Henry John Todd in 1818 and by Robert Gordon Latham in 1866.

It was unfortunate that Joseph Priestley, Robert Lowth, James Buchanan, and other 18th-century grammarians (Priestley was perhaps better known as a scientist and theologian) took a narrower view than Johnson on linguistic growth and development. They spent too much time condemning such current "improprieties" as "I had rather not," "you better go," "between you and I," "it is me," "who is this for?", "between four walls," "a third alternative," "the largest of the two," "more perfect," and "quite unique." Without explanatory comment they banned "you was" outright, although it was in widespread use among educated people (on that ground it was later defended by Noah Webster). "You was" had, in fact, taken the place of both "thou wast" and "thou wert" as a useful singular equivalent of the accepted plural "you were."

As the century wore on, grammarians became more numerous and aggressive. They set themselves up as arbiters of correct usage. They compiled manuals that were not only descriptive (stating what people do say) and prescriptive (stating what they should say) but also proscriptive (stating what they should not say). They regarded Latin as a language superior to English and claimed that Latin embodied universally valid canons of logic. This view was well maintained by Lindley Murray, a native of Pennsylvania who settled in England in the very year (1784) of Johnson's death. Murray's *English Grammar* appeared in 1795, became immensely popular, and went into numerous editions. It was followed by an *English Reader* (1799) and an *English Spelling Book* (1804), long favourite textbooks in both Old and New England.

Proposals for a new dictionary

19th and 20th centuries. In 1857 Richard Chenevix Trench, dean of St. Paul's, lectured to the Philological Society on the theme, "On some Deficiencies in our English Dictionaries." His proposals for a new dictionary were implemented in 1859, when Samuel Taylor Coleridge's grandnephew, Herbert Coleridge, set to work as first editor. He was succeeded by a lawyer named Frederick James Furnivall, who in 1864 founded the Early English Text Society with a view to making all the earlier literature available to historical lexicographers in competent editions. Furnivall was subsequently succeeded as editor by James A.H. Murray, who published the first fascicle of *A New English Dictionary on Historical Principles* in 1884. Later Murray was joined successively by three editors: Henry Bradley, William Alexander Craigie, and Charles Talbut Onions. Aside from its *Supplements*, the dictionary itself fills 12 volumes, has over 15,000 pages, and contains 414,825 words, illustrated by 1,827,306 citations. It is a dictionary of the British Commonwealth and the United States, a fact symbolized by the presentation of first copies in the spring of 1928 to King George V and Pres. Calvin Coolidge. It exhibits the histories and meanings of all words known to have been in use since 1150. From 1150 to 1500 all five Middle English dialects, as has been seen, were of equal status. They are therefore all included. After 1500, however, dialectal expressions are not admitted, nor are scientific and technical terms not in general use. Otherwise, the written vocabulary is comprehensive. A revised edition of this dictionary, known as *The Oxford English Dictionary*, was published in 1933.

VARIETIES OF ENGLISH

British English. The abbreviation RP (Received Pronunciation) denotes the speech of educated people living in London and the southeast of England and of other people elsewhere who speak in this way. If the qualifier educated be assumed, RP is then a regional (geographical) dialect, as contrasted with London Cockney, which is a class (social) dialect. RP is not intrinsically superior to other varieties of English; it is itself only one particular regional dialect that has, through the accidents of history, achieved more extensive use than others. Although acquiring its unique status without the aid of any established authority, it may have been fostered by the public schools (Winchester, Eton, Harrow, Rugby, and so on) and the ancient universities (Oxford and Cambridge). Other varieties of English are well preserved in spite of the levelling influences of film, television, and radio. In the Northern dialect RP /ɑː/ (the first vowel sound in "father") is still pronounced /æ/ (a sound like the *a* in "fat") in words such as laugh, fast, and path; this pronunciation has been carried across the Atlantic into American English.

Features of the Northern dialect

In the words run, rung, and tongue, the received-standard pronunciation of the vowel is /ʌ/, like the *u* in "but"; in the Northern dialect it is /u/, like the *oo* in "book." In the words bind, find, and grind, the received standard pronunciation of the vowel sound is /ai/, like that in "bide"; in Northern, it is /i/, like the sound in "feet." The vowel sound in the words go, home, and know in the Northern dialect is /ɔː/, approximately the sound in "law" in some American English dialects. In parts of Northumberland, RP "it" is still pronounced "hit," as in Old English. In various Northern dialects the definite article "the" is heard as *t*, *th*, or *d*. In those dialects in which it becomes both *t* and *th*, *t* is used before consonants and *th* before vowels. Thus, one hears "t'book" but "th'apple." When, however, the definite article is reduced to *t* and the following word begins with *t* or *d*, as in "t'tail" or "t'dog," it is replaced by a slight pause as in the RP articulation of the first *t* in "hat trick." The RP /tʃ/, the sound of the *ch* in "church," becomes *k*, as in "thack," ("thatch, roof") and "kirk" ("church"). In many Northern dialects strong verbs retain the old past-tense singular forms band, brak, fand, spak for RP forms bound, broke, found, and spoke. Strong verbs also retain the past participle inflection *-en* as in "comen," "shutten," "sitten," and "getten" or "gotten" for RP "come," "shut," "sat," and "got."

In some Midland dialects the diphthongs in "throat" and "stone" have been kept apart, whereas in RP they have fallen together. In Cheshire, Derby, Stafford, and Warwick, RP "singing" is pronounced with a *g* sounded after the velar nasal sound (as in RP "finger"). In Norfolk one hears "skellington" and "solintary" for "skeleton" and "solitary," showing an intrusive *n* just as does "messenger" in RP from French *messager*, "passenger" from French *passager*, and "nightingale" from Old English *nihtegala*. Other East Anglian words show consonantal metathesis (switch position), as in "singify," and substitution of one liquid or nasal for another, as in "chimbly" for "chimney," and "synnable" for "syllable." "Hantle" for "handful" shows syncope (disappearance) of an unstressed vowel, partial assimilation of *d* to *t* before voiceless *f*, and subsequent loss of *f* in a triple consonant group.

South Western dialects

In South Western dialects, initial *f* and *s* are often voiced, becoming *v* and *z*. Two words with initial *v* have found their way into RP: "vat" from "fat" and "vixen" from "fixen" (female fox). Another South Western feature is the development of a *d* between *l* or *n* and *r*, as in "parlder" for "parlour" and "carnder" for "corner." The bilabial semivowel *w* has developed before *o* in "wold" for "old," and in "wom" for "home," illustrating a similar development in RP by which Old English *ān* has become "one," and Old English *hāl* has come to be spelled "whole," as compared with Northern *hale*. In South Western dialects "yat" comes from the old singular *geat*, whereas RP "gate" comes from the plural *gatu*. Likewise, "clee" comes from the old nominative *clea*, whereas RP "claw" comes from the oblique cases. The verbs keel and kemb have developed regularly from Old English *cēlan* "to make cool" and *kemban* "to use a comb," whereas the corresponding RP verbs cool and comb come from the adjective and the noun, respectively.

In Wales, people often speak a clear and measured form of English with a musical intonation inherited from ancestral Celtic. They tend to aspirate both plosives (stops) and fricative consonants very forcibly; thus, "true" is pronounced with an audible puff of breath after the initial *t*.

Lowland Scottish was once a part of Northern English, but two dialects began to diverge in the 14th century. Today Lowland Scots trill their *r*'s, shorten vowels, and simplify diphthongs. A few Scottish words, such as bairn, brae, canny, dour, and pawky, have made their way into RP. Lowland Scottish is not to be confused with Scottish Gaelic, a Celtic language still spoken by about 90,700 people (almost all bilingual) mostly in the Highlands and the Western Isles. Thanks to Robert Burns and Sir Walter Scott, many Scottish Gaelic words have been preserved in English literature.

Irish pronunciation

Northern Ireland has dialects related in part to Lowland Scottish and in part to the southern Irish dialect of English. Irish pronunciation is conservative and is clearer and more easily intelligible than many other dialects. Its literature has reached worldwide audiences, whether written by Englishmen born in Ireland, such as Jonathan Swift, Laurence Sterne, Oliver Goldsmith, Sir Richard Steele, Edmund Burke, Oscar Wilde, and George Bernard Shaw, or by authentic Irish, such as James Joyce, William Butler Yeats, and John Millington Synge. The influence of Irish Gaelic on the speech of Dublin is most evident in the syntax of drama and in the survival of such picturesque expressions as "We are after finishing," "It's sorry you will be," and "James do be cutting corn every day."

American and Canadian English. The dialect regions of the United States are most clearly marked along the Atlantic littoral, where the earlier settlements were made. Three dialects can be defined: Northern, Midland, and Southern. Each has its subdialects.

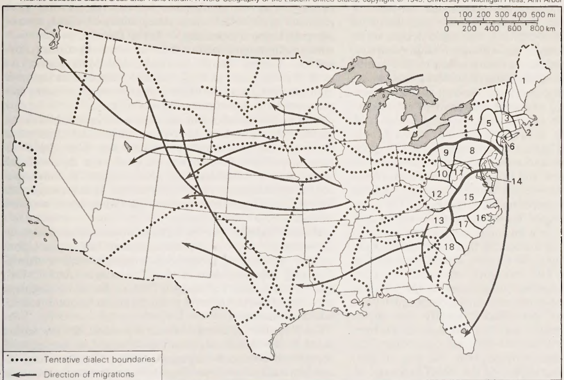
The Northern dialect is spoken in New England. Its six chief subdialects comprise northeastern New England (Maine, New Hampshire, and eastern Vermont), southeastern New England (eastern Massachusetts, eastern Connecticut, and Rhode Island), southwestern New England (western Massachusetts and western Connecticut), the inland north (western Vermont and upstate New York), the Hudson Valley, and metropolitan New York (Figure 15).

The Midland dialect is spoken in the coastal region from Point Pleasant, in New Jersey, to Dover, in Delaware. Its seven major subdialects comprise the Delaware Valley, the Susquehanna Valley, the Upper Ohio Valley, northern West Virginia, the Upper Potomac and Shenandoah, southern West Virginia and eastern Kentucky, western Carolina, and eastern Tennessee.

The Southern dialect area covers the coastal region from Delaware to South Carolina. Its five chief subdialects comprise the Delmarva Peninsula, the Virginia Piedmont, northeastern North Carolina (Albemarle Sound and Neuse Valley), Cape Fear and Pee Dee valleys, and the South Carolina Low Country, around Charleston.

These boundaries, based on those of the *Linguistic Atlas of the United States and Canada*, are highly tentative. To some extent these regions preserve the traditional speech of southeastern and southern England, where most of the early colonists were born. The first settlers who came to Virginia (1607) and Massachusetts (1620) soon learned to adapt old words to new uses, but they were content to borrow names from the local Indian languages for unknown trees, such as hickory and persimmon, and for unfamiliar animals, such as raccoons and woodchucks. Later they took words from foreign settlers: "chowder" and "prairie" from the French, "scow" and "sleigh" from the Dutch. They made new compounds, such as "backwoods" and "bullfrog," and gave new meanings to such words as "lumber" (which in British English denotes disused furniture,

Prepared by Mrs. Raven I. McDavid, Jr., for *The Structure of American English*, by W. Nelson Francis, copyright © 1958, The Ronald Press Company, New York. Atlantic Seaboard dialect areas after Hans Kurath, *A Word Geography of the Eastern United States*; copyright © 1949, University of Michigan Press, Ann Arbor



ATLANTIC SEABOARD DIALECT AREAS

THE NORTH
1. Northeastern New England
2. Southeastern New England
3. Southwestern New England
4. Inland North (western Vermont, Upstate New York and derivatives)
5. The Hudson Valley
6. Metropolitan New York

THE MIDLAND (North)
7. Delaware Valley (Philadelphia)
8. Susquehanna Valley
9. Upper Ohio Valley (Pittsburgh)
10. Northern West Virginia

THE MIDLAND (South)
11. Upper Potomac and Shenandoah
12. Southern West Virginia and Eastern Kentucky
13. Western Carolina and Eastern Tennessee

THE SOUTH
14. Delmarva (Eastern Shore)
15. The Virginia Piedmont
16. Northeastern North Carolina (Albemarle Sound and Neuse Valley)
17. Cape Fear and Peedee Valleys
18. The South Carolina Low Country (Charleston)

Figure 15: Dialect areas of English in the United States.

or junk) and "corn" (which in British English signifies any grain, especially wheat).

Before the Declaration of Independence (1776), two-thirds of the immigrants had come from England, but after that date they arrived in large numbers from Ireland. The potato famine of 1845 drove 1,500,000 Irish to seek homes in the New World, and the European revolutions of 1848 drove as many Germans to settle in Pennsylvania and the Middle West. After the close of the American Civil War, millions of Scandinavians, Slavs, and Italians crossed the ocean and eventually settled mostly in the North Central and Upper Midwest states. In some areas of South Carolina and Georgia the American Negroes who had been imported to work the rice and cotton plantations developed a contact language called Gullah, or Geechee, that made use of many structural and lexical features of their native languages. This remarkable variety of English is comparable to such "contact languages" as Sranan (Taki-Taki) and Melanesian Pidgin. The speech of the Atlantic Seaboard shows far greater differences in pronunciation, grammar, and vocabulary than that of any area in the North Central States, the Upper Midwest, the Rocky Mountains, or the Pacific Coast. Today, urbanization, quick transport, and television have tended to level out some dialectal differences in the United States.

The Gullah dialect

The boundary with Canada nowhere corresponds to any boundary between dialects, and the influence of United States English is strong, being felt least in the Maritime Provinces and Newfoundland. Nevertheless, in spite of the effect of this proximity to the United States, British influences are still potent in some of the larger cities; Scottish influences are well sustained in Ontario. Canada remains bilingual. One-fourth of its people, living mostly in the province of Quebec, have French as their mother tongue. Those provinces in which French is spoken as a mother tongue by 10 percent or more of the population are described as "federal bilingual districts" in the Official Languages Bill of 1968.

Australian and New Zealand English. Unlike Canada, Australia has few speakers of European languages other than English within its borders. There are still many Aboriginal languages, though they are spoken by only a few hundred speakers each and their continued existence is threatened. More than 80 percent of the population is British. By the mid-20th century, with rapid decline of its Aboriginal tongues, English was without rivals in Australia.

During colonial times the new settlers had to find names for a fauna and flora (*e.g.*, banksia, iron bark, whee whee) different from anything previously known to them: trees that shed bark instead of leaves and cherries with external stones. The words brush, bush, creek, paddock, and scrub acquired wider senses, whereas the terms brook, dale, field, forest, and meadow were seldom used. A creek leading out of a river and entering it again downstream was called an anastomizing branch (a term from anatomy), or an anabranch, whereas a creek coming to a dead end was called by its native name, a billabong. The giant kingfisher with its raucous bray was long referred to as a laughing jackass, later as a bushman's clock, but now it is a kookaburra. Cattle so intractable that only roping could control them were said to be ropable, a term now used as a synonym for "angry" or "extremely annoyed."

A deadbeat was a penniless "sundowner" at the very end of his tether, and a no-hoper was an incompetent fellow, hopeless and helpless. An offsider (strictly, the offside driver of a bullock team) was any assistant or partner. A rouseabout was first an odd-job man on a sheep station and then any kind of handyman. He was, in fact, the "down-under" counterpart of the wharf labourer, or roustabout, on the Mississippi River. Both words originated in Cornwall, and many other terms, now exclusively Australian, came ultimately from British dialects. "Dinkum," for instance, meaning "true, authentic, genuine," echoed the "fair dinkum," or fair deal, of Lincolnshire dialect. "Fossicking" about for surface gold, and then rummaging about in general, perpetuated the term fossick ("to elicit information, ferret out the facts") from the Cornish dialect of English. To "barrack," or jeer noisily, recalled Irish "barrack" ("to brag, boast"), whereas "skerrick" in the phrase "not a skerrick left" was obviously identical with the "skerrick" meaning "small fragment, particle," still heard in English dialects from Westmorland to Hampshire.

"Rouseabout" and "roustabout"

Some Australian English terms came from Aboriginal speech: the words boomerang, corroboree (warlike dance and then any large and noisy gathering), dingo (reddish-brown half-domesticated dog), galah (cockatoo), gunyah (bush hut), kangaroo, karri (dark-red eucalyptus tree), nonda (rosaceous tree yielding edible fruit), pokutukawa (evergreen bearing brilliant blossom), wallaby (small marsupial), and wallaroo (large rock kangaroo). Australian English has slower rhythms and flatter intonations than RP. Although there is remarkably little regional variation throughout the entire continent, there is significant social variation. The neutral vowel /ə/ (as the *a* in "sofa") is frequently used, as in London Cockney: "arches" and "archers" are both pronounced [ɑːtʃəz], and the pronunciations of RP "day" and "go" are, respectively, [dəi] and [gəu].

Although New Zealand lies over 1,000 miles away, much of the English spoken there is similar to that of Australia. The blanket term Austral English is sometimes used to cover the language of the whole of Australasia, or Southern Asia, but this term is far from popular with New Zealanders because it makes no reference to New Zealand and gives all the prominence, so they feel, to Australia. Between North and South Islands there are observable differences. For one thing, Maori, which is still a living language (related to Tahitian, Hawaiian, and the other Austronesian [Malayo-Polynesian] languages), has a greater number of speakers and more influence in North Island.

The English of India–Pakistan. In 1950 India became a federal republic within the Commonwealth of Nations, and Hindi was declared the first national language. English, it was stated, would "continue to be used for all official purposes until 1965." In 1967, however, by the terms of the English Language Amendment Bill, English was proclaimed "an alternative official or associate language with Hindi until such time as all non-Hindi states had agreed to its being dropped." English is therefore acknowledged to be indispensable. It is the only practicable means of day-to-day communication between the central government at New Delhi and states with non-Hindi speaking populations, especially with the Deccan, or "South," where millions speak Dravidian (non-Indo-European) languages—Telugu, Tamil, Kannada, and Malayalam. English is widely used in business, and, although its use as a medium in higher education is decreasing, it remains the principal language of scientific research.

English an alternative official language

In 1956 Pakistan became an autonomous republic comprising two states, East and West. Bengali and Urdu were made the national languages of East and West Pakistan, respectively, but English was adopted as a third official language and functioned as the medium of interstate communication. (In 1971 East Pakistan broke away from its western partner and became the independent state of Bangladesh.)

African English. Africa is the most multilingual area in the world, if people are measured against languages. Upon a large number of indigenous languages rests a slowly changing superstructure of world languages (Arabic, English, French, and Portuguese). The problems of language are everywhere linked with political, social, economic, and educational factors.

The Republic of South Africa, the oldest British settlement in the continent, resembles Canada in having two recognized European languages within its borders: English and Afrikaans, or Cape Dutch. Both British and Dutch traders followed in the wake of 15th-century Portuguese explorers and have lived in widely varying war-and-peace relationships ever since. Although the Union of South Africa, comprising Cape Province, Transvaal, Natal, and Orange Free State, was for more than a half century (1910–61) a member of the British Empire and Commonwealth, its four prime ministers (Botha, Smuts, Hertzog, and Malan) were all Dutchmen. In the early 1980s Afrikaners outnumbered Britishers by three to two. The Afrikaans language began to diverge seriously from European Dutch in the late 18th century and has gradually come to be recognized as a separate language. Although the English spoken in South Africa differs in some respects from standard British English, its speakers do not regard the language as a separate one. They have naturally come to use many Afrikanerisms, such as *kloof, kopje, krans, veld,* and *vlei,* to denote features of the landscape and occasionally employ African names to designate local animals and plants. The words trek and commando, notorious in South African history, have acquired almost worldwide currency.

Afrikanerisms in South African English

Elsewhere in Africa, English helps to answer the needs of wider communication. It functions as an official language of administration in Botswana, Lesotho, and Swaziland and in Zimbabwe, Zambia, Malaŵi, Uganda, and Kenya. It is the language of instruction at Makerere University in Kampala, Uganda; at the University of Nairobi, Kenya; and at the University of Dar es Salaam, in Tanzania.

The West African states of The Gambia, Sierra Leone, Ghana, and Nigeria, independent members of the Commonwealth, have English as their official language. They are all multilingual. The official language of Liberia is also English, although its tribal communities constitute four different linguistic groups. Its leading citizens regard themselves as Americo-Liberians, being descendants of those freed blacks whose first contingents arrived in West Africa in 1822. South of the Sahara indigenous languages are extending their domains and are competing healthily and vigorously with French and English.

THE FUTURE OF ENGLISH

Geographically, English is the most widespread language on earth, and it is second only to Mandarin Chinese in the number of people who speak it. The International Telecommunication Union (ITU) has five official languages: English, French, Spanish, Russian, and Chinese. The influence of these languages upon one another will inevitably increase.

It is reasonable to ask if changes in English can be predicted. There will doubtless be modifications in pronunciation, especially in that of long vowels and diphthongs. In weakly stressed syllables there is already a discernible tendency, operating effectively through radio and television, to restore the full qualities of vowels in these syllables. This tendency may bring British English more into line with American English and may bring them both a little nearer to Spanish and Italian. Further, it may help to narrow the gap between pronunciation and spelling. Other factors will also contribute toward the narrowing of this gap: advanced technological education, computer programming, machine translation, and expanding mass media. Spelling reformers will arise from time to time to liven up proceedings, but, in general, traditional orthography may well hold its own against all comers, perhaps with some regularization. Printing houses, wielding concentrated power through their style directives, will surely find it in their best interests to agree on uniformity of spelling. Encyclopaedic dictionaries—computerized, universal, and subject to continuous revision—may not go on indefinitely recording such variant spellings as "connection" and "connexion," "judgment" and "judgement," "labor" and "labour," "medieval" and "mediaeval," "plow" and "plough," "realise" and "realize," "thru" and "through."

Factors leading toward conformity

Since Tudor days, aside from the verb endings *-est* and *-eth,* inflections have remained stable because they represent the essential minimum. The abandonment of the forms thou and thee may encourage the spread of yous and youse in many areas, but it is not necessarily certain that these forms will win general acceptance. The need for a distinctive plural can be supplied in other ways (*e.g.,* the forms "you all, you fellows, you people"). The distinctions between the words "I" and "me," "he" and "him," "she" and "her," "we" and "us," "they" and "them" seem to many authors to be too important to be set aside, in spite of a growing tendency to use objective forms as emphatic subjective pronouns and to say, for instance, "them and us" instead of "they and we" in contrasting social classes. Otherwise, these distinctive forms may remain stable; they are all monosyllabic, they are in daily use, and they can bear the main stress. Thus they are likely to resist levelling processes.

Considerable changes will continue to be made in the forms and functions of auxiliary verbs, catenative (linking) verbs, phrasal verbs, and verb phrases. Indeed, the constituents of verbs and verb groups are being more subtly modified than those of any other word class. By means of auxiliaries and participles, a highly intricate system of aspects, tenses, and modalities is gradually evolving.

In syntax the movement toward a stricter word order seems to many to be certain to continue. The extension of multiple attributives in nominal groups has probably reached its maximum. It cannot extend further without incurring the risk of ambiguity.

In vocabulary further increases are expected if the present trends continue. Unabbreviated general dictionaries already contain 500,000 entries, but even larger dictionaries, with 750,000 entries, may be required. Coiners of words probably will not confine themselves to Greek and Latin in creating new terms; instead they are likely to exercise their inventive powers in developing an international technical vocabulary that is increasingly shared by Russian, French, and Spanish and that is slowly emerging as the universal scientific language. (S.P.)

Armenian language

The Armenian language, which forms a separate branch of the western group of Indo-European languages, is the mother tongue of the Turkish Armenians and of the Armenians in Armenia, where it is spoken by 2,850,000 people. In other parts of the former Soviet Union, especially in the neighbouring republics of Georgia and Azerbaijan, it is used by some 1,300,000. Armenian emigrants and refugees have taken their language with them all over Asia Minor and the Middle East and from there to many European countries, especially Romania, Poland, and France, and to America, particularly the United States. In all, Armenian is probably spoken by about 5,500,000 people around the world.

History of the language. Armenian was introduced into the mountainous Transcaucasian region (called Greater Armenia by the Greek historians) by invaders coming from the northern Balkans, probably in the latter part of the 2nd millennium BC. These invaders occupied the region on the shores of Lake Van that had previously been the site of the ancient Urartean kingdom. By the 7th century BC the Armenian language seems to have re-

Table 28: The Armenian Alphabet

letter		equivalent	letter		equivalent
capital	lowercase		capital	lowercase	
Ա	ա	a	Մ	մ	m
Բ	բ	b	Յ	յ	y
Գ	գ	g	Ն	ն	n
Դ	դ	d	Շ	շ	sh
Ե	ե	e	Ո	ո	o
Զ	զ	z	Չ	չ	ch'*
Է	է	ē	Պ	պ	p
Ը	ը	ĕ	Ջ	ջ	j
Թ	թ	t'*	Ռ	ռ	rh
Ժ	ժ	zh	Ս	ս	s
Ի	ի	i	Վ	վ	v
Լ	լ	l	Տ	տ	t
Խ	խ	kh	Ր	ր	r
Ծ	ծ	ts	Ց	ց	ts'*
Կ	կ	k	Ւ	ւ	w
Հ	հ	h	Փ	փ	p'*
Ձ	ձ	dz	Ք	ք	k'*
Ղ	ղ	gh	Օ	օ	ō
Ճ	ճ	ch	Ֆ	ֆ	f

*The *spiritus asper,* ', indicates aspiration.

placed the tongues of the native population. It is tempting to connect the invasion with the downfall of the Hittite empire in Anatolia.

Development of the alphabet

After the introduction of Christianity to Armenia about AD 400, the language began to be written down; an alphabet of 36 letters was invented, according to tradition, by Mesrop Mashtots (Table 28). (Two letters were added later.) Admirably suited to the phonology of Armenian, it is still used in various forms by Armenians all over the world. The oldest writings in the language date from the 5th century; they are preserved in manuscript form from the 9th century. Grabar, the written language of the 5th century, the golden age of Armenian culture, is traditionally said to be based on the dialect of Tarawn on Lake Van. To what extent the spoken language was split into dialects at that time is not known. The language of the literature from the 5th to the 8th century is remarkably homogeneous, but by the 9th century the influence of the spoken dialects was noticeable, especially in legal and historical texts. Among the Middle Armenian varieties of Grabar, the best known is the 12th- and 13th-century chancellery (court) language of the Armenian kingdom in Cilicia. More or less corrupted versions of Grabar continued as the literary language until the middle of the 19th century.

Modern East Armenian and West Armenian. In the 1800s, the writers Khachatur Abovean (1805–48) and Mikael Nalbandean (1829–66) and other Armenian nationalists made efforts to reach the populace with nationalist propaganda. As a result, a national revival occurred from which a new literary language emerged that was much closer to the spoken language. This is known in two varieties. East Armenian, now the official language of Armenia, is based on the dialect of the Ararat valley and the city of Yerevan; West Armenian has its foundation in the dialect of Istanbul. East Armenian is also spoken in other parts of Transcaucasia, whereas the western variety dominates in the Armenian colonies in the Middle East, Asia Minor, Europe, and America. The differences between these two written forms of Modern Armenian are slight, constituting no barrier to mutual intelligibility.

Dialects

In addition to the two literary languages, there are a great number of dialects, some of which are so different that the speakers cannot understand each other. It is estimated that before World War I some 50 distinct dialects were spoken. These dialects, however, have lost ground in Transcaucasia, under the pressure of the standard written language.

When the scientific study of Armenian started in the 19th century, the language was considered an Iranian dialect, a mistake easily explained by the vast number of Iranian loanwords in the vocabulary. Subsequent studies, however, have convincingly shown Armenian to be an independent member of the Indo-European language family. According to the Greek historian Herodotus, Armenian was a variety of Phrygian, a tongue presumed to be Indo-European. What little is known of the latter is insufficient to support or confirm such a claim.

Sounds of Armenian

Phonology. Phonetic developments in Armenian have radically changed the sound system of the old Indo-European parent language. In particular, the pattern of the plosive consonants—the stops—has been completely reshuffled. In the more conservative central Armenian dialects, three series of stops are distinguished (voiced *b, d, g,* which in some dialects are aspirated; unvoiced *p, t, k;* and unvoiced aspirated *ph, th, kh*). In the dialects of the periphery, the three series have been reduced to two (aspirated *ph, th, kh* and unaspirated *p, t, k* or, as in Istanbul, *b, d, g*). These differences are concealed in the traditional orthography. Sibilants (fricatives) of various types and affricates have emerged through palatalization of the old palatal and labiovelar stops. Thus, Old Armenian *dz* (*z*) and *ts* may go back to the Indo-European palatal stops *ǵh* and *ǵ*, and *dž* (*ž*) to the Indo-European labiovelar g^wh before *e* and *i;* Old Armenian *ts*ʿ, *tš*ʿ, and *tš* may derive from the Indo-European consonant clusters *sk, ky,* and *gy.* All Armenian dialects distinguish two types of *r,* one strongly trilled, one weakly trilled. Old Armenian also differentiated between two types of *l,* neutral and velarized, the latter of which is made by moving the back of the tongue nearer to the soft palate at the back of the mouth. The latter type has developed into a voiced velar fricative in the modern dialects.

Grammar. All the spoken dialects and the two literary languages have maintained a fairly complicated system of noun declension, distinguishing six or seven cases. The plural stem, derived from the singular stem by the addition of the suffix -(*n*)*er,* is declined as a singular, according to the Turkish pattern. Characteristic of the changes in the Old Armenian verbal system is the general replacement of simple present tense forms by periphrastic expressions. These are groups of words, including auxiliaries, that take the place of a single word that is capable of being inflected to show tense or some other feature. The various types of periphrastic forms serve as the basis for the classification of the dialects. In Old and Modern Armenian, the main tense distinction is that between present, aorist (denoting occurrence without reference to completeness, duration, and so on), and periphrastic perfect tenses. The old subjunctive, still extant in classical Armenian, has been lost in the modern language. To express future time, Old Armenian used the subjunctive forms; Modern Armenian employs periphrastic expressions.

Also characteristic of Modern Armenian is the importance of the passive forms of the verb, which are strictly parallel to the active forms, and the emergence of a special negative conjugation with differing forms for verbs in instances like *I read* and *I don't read.* Whereas Old Armenian was rather close to ancient Greek in many respects, Modern Armenian is typologically much closer to Turkish. Among the features that illustrate this similarity are the agglutinative system of declension (that is, the compounding of several linguistic elements of independent meaning into a single word), the use of suffixes to indicate possession, the employing of passive and causative forms for all verbs, and the use of postpositions (grammatical elements that are placed after the word) instead of prepositions (as were used in Old Armenian). The vocabulary of the written languages is purely Armenian, being based almost exclusively on that of Grabar, with very few loanwords from the neighbouring languages. (The Iranian loanwords mentioned above were incorporated into Armenian before the creation of the written language.) (H.K.V./Ed.)

Tocharian languages

Tocharian (Tokharian) is a small group of Indo-European languages that was spoken in the Tarim Basin (in the centre of what is now the Uighur Autonomous Region of Sinkiang, China) during the latter half of the 1st millennium AD. Documents from AD 500–700 attest to two languages: Tocharian A, from the area of Turfan in the east; and Tocharian B, chiefly from the region of Kucha in the west but also from the Turfan area.

HISTORY AND LINGUISTIC CHARACTERISTICS

Discovery and decipherment. The first Tocharian manuscripts were discovered in the 1890s. The bulk of the Tocharian materials were carried to Berlin by the Prussian expeditions of 1903–04 and 1906–07, which explored the Turfan area, and to Paris by a French expedition of 1906–09, which investigated chiefly in the area of Kucha. Smaller collections are in London, Calcutta, St. Petersburg, and Japan, the result of Indo-British, Russian, and Japanese expeditions.

Translation and analysis

The Tocharian languages are written in a northern Indian syllabary (a set of characters representing syllables) known as Brāhmī, which was also used in writing Sanskrit manuscripts from the same area. The first successful attempt at grammatical analysis and translation was made by the German scholars Emil Sieg and Wilhelm Siegling in 1908 in an article that also established the presence of the two languages (sometimes referred to as dialects), provisionally called A and B. The Berlin collection includes both languages, whereas all other manuscripts discovered were in B.

The German name Tocharisch was proposed (see below *The "Tocharian problem"*), and the language was demonstrated to be Indo-European.

Literature. Tocharian literature is Buddhistic in content, consisting largely of translations or free adaptations of Jātakas, of Avadānas, and of philosophical, didactic, and canonical works. In Tocharian B there are also commercial documents, such as monastery records, caravan passes, medical and magical texts, and the like. These are important source materials for information on the social, economic, and political life of Central Asia.

Linguistic characteristics. Tocharian forms an independent branch of the Indo-European language family not closely related to other neighbouring Indo-European languages (Indo-Aryan and Iranian). Rather, Tocharian shows a closer affinity with the western (centum) languages: compare, for example, Tocharian A *känt,* B *kante* '100' and Latin *centum* with Sanskrit *śatám*; A *klyos-,* B *klyaus-* 'hear' and Latin *clueo* with Sanskrit *śru-*; A *kus,* B *kuse* 'who' and Latin *qui, quod* with Sanskrit *kas.* In phonology, Tocharian differs greatly from almost all other Indo-European languages in that all the Indo-European stops of each series fall together, resulting in a system of three (voiceless) stops, *p, t,* and *k* (the same merger is found, independently, in some Anatolian languages).

The Tocharian verb reflects the Indo-European verbal system both in stem formations and in personal endings. Especially noteworthy is the wide development of the mediopassive form in *r* (as in Italic and Celtic)—*e.g.,* Tocharian A *klyoṣtär,* B *klyaustär* 'is heard.' The third person plural preterite (past) ends in *-r,* similar to Latin and Sanskrit perfect forms and the Hittite preterite. The noun shows less of its Indo-European origins. However, it preserves three numbers (singular, dual, and plural) and traces at least of the nominative, accusative, genitive, vocative, and ablative cases. Most of the attested cases are built up by the addition of postpositions to the oblique (accusative) form.

The vocabulary shows the influence of Iranian and, later, Sanskrit (the latter language particularly was the source of Buddhist terminology). Chinese had little influence (a few weights and measures and the name of at least one month). Many of the most archaic elements of the Indo-European vocabulary are retained—*e.g.,* A *por,* B *puwar* 'fire' (Greek *pyr,* Hittite *paḫḫur*); A and B *ku* 'dog' (Greek *kyōn*); A *tkaṃ,* B *keṃ* 'earth' (Greek *chthōn,* Hittite *tekan*); and, especially, nouns of relationship: A *pācar, mācar, pracar, ckācar,* B *pācer, mācer, procer, tkācer,* 'father,' 'mother,' 'brother,' and 'daughter,' respectively.

THE "TOCHARIAN PROBLEM"

Since the appearance of Sieg and Siegling's article, the appropriateness of the name Tocharian for these languages has been disputed. Their use of the word Tocharisch is largely based on the statement in a copy of a Buddhist drama written in a form of Turkish that it had been translated from "Twgry," and, because the work is otherwise known only in Tocharian A, it was natural to make the equation of Twgry with Tocharian A. The equation of the Twgry language with that of the Tocharoi is based on phonetic similarity. According to Greek and Latin historical sources, the Tocharoi (Greek Tócharoi, Latin Tochari, Sanskrit *tukhāra*) inhabited the basin of the upper Oxus River (modern Amu Darya) in the 2nd century BC, having been driven there from an earlier home in Kansu (immediately east of Sinkiang).

In later times their ruling elite used a form of Iranian as a written language but what their original language may have been remains uncertain. A Sanskrit–Tocharian B bilingual inscription appears to equate Sanskrit *tokharika* and Tocharian B *kucaññe* 'Kuchean.' However, the rest of the context remains obscure. Thus, Sieg and Siegling's identification of the Tocharian language family and the classical Tocharoi remains most speculative, though the designation Tocharian seems fixed nonetheless. For language A and language B, the substitution of Turfanian and Kuchean, or of East Tocharian and West Tocharian, is sometimes found.

Despite its historical position on the eastern frontier of the Indo-European world and the obvious lexical influence of Indo-Aryan and Iranian, Tocharian seems more closely allied linguistically with languages of the Indo-European northwest, particularly Italic and Germanic, in the matter of common vocabulary and certain verbal categories. To a lesser extent Tocharian appears to share certain features with Balto-Slavic and Greek.

With regard to the two Tocharian languages themselves, it is possible that Tocharian A was, at the time of documentation, a dead liturgical language preserved in the Buddhist monasteries in the east, whereas Tocharian B was a living language in the west. The presence of manuscripts in B mixed with those in A in the monasteries of the east can be accounted for by ascribing the B manuscripts to a new missionary initiative by Buddhist monks from the west. (G.S.L./D.Q.A.)

Celtic languages

The Celtic languages form one branch of the Indo-European language family, having in common certain sound shifts and vocabulary items. On both geographic and chronological grounds, the languages fall into two divisions, usually known as Continental Celtic and Insular Celtic.

Continental Celtic. Continental Celtic is the generic name for the languages spoken by the people known to classical writers as Keltoi and Galatae; at various times during a period of roughly 1,000 years (approximately 500 BC–AD 500), they occupied an area that stretched from Gaul to Iberia in the south and Galatia in the east. The great bulk of evidence for Continental Celtic consists of the names of persons, tribes, and places recorded by Greek and Latin writers. Only in Gaul and in northern Italy are inscriptions found, and the interpretation of these is in most cases doubtful. Given the nature of the evidence, knowledge of these languages is confined largely to the sound system and a small part of the vocabulary, and no certain conclusions can be reached as to their historical development or the differences between them.

Evidence for Continental Celtic

Insular Celtic. Insular Celtic refers to the Celtic languages of the British Isles, together with Breton (spoken in Brittany, France). As the name Breton implies, it is an importation from Britain and is not a Continental Celtic dialect. Although there is some scanty evidence from classical sources—mainly place-names—and a small body of inscriptions in the Latin and ogham alphabets from the end of the 4th to the 8th century AD, the main sources of information on the early stages of these languages are manuscripts written from the 7th century onward in Irish and somewhat later in the British languages.

The Insular languages fall into two groups—Irish and British. Irish (often called Goidelic, from Old Irish *Goídel* "Irishman," or Gaelic, from *Gael,* the modern form of the same word) was the only language spoken in Ireland in the 5th century, the time when historical knowledge of that island begins. The two other members of this group, Scottish Gaelic and Manx, arose from Irish colonizations that began about that time. There were also important Irish-speaking colonies in Wales, but no trace of their language survives apart from a few inscriptions.

British (often called Brythonic, from Welsh Brython "Briton") had almost the same degree of influence on the island of Britain and the Isle of Man. Inscriptions and personal names surviving from Scotland show clearly that there was a non-Indo-European language spoken there, usually called Pictish, which was later replaced by British. There were undoubtedly dialectal differences within the island, but the existing dialects arose from the fragmentation of British by the Irish invasions of Man and what is now Scotland and by the English invasions that began in what is now southern England and finally reached Scotland. Scotland has ever since been partitioned linguistically between English (or "Scots") and Irish (or "Erse"—the Scots form of "Irish"—or "Gaelic"). A British dialect, now labeled Cumbric, lingered on in the western borderlands between England and Scotland until perhaps the 10th century, but almost nothing is known about it. In what is now Wales, British survived as the dominant language until a century or so ago; it is now known as Welsh. Another pocket of British speech survived in Cornwall until the end of the 18th century. It was from this area that emigrants in the

Re-establishment of Celtic on the European mainland

5th and 6th centuries AD had brought Celtic once more to the European mainland by establishing a colony in northwestern France, still called Brittany. It is just possible that there were some traces of the Continental Celtic language (*i.e.*, Gaulish) at that time in this remote area, although Breton is too similar to Cornish (an Insular Celtic tongue) to suggest any serious influence from Gaulish.

HISTORICAL DEVELOPMENT

Common Celtic. The reconstruction of Common Celtic (or Proto-Celtic)—the parent language that yielded the various tongues of Continental Celtic and Insular Celtic—is of necessity very tentative. Whereas Continental Celtic offers plenty of evidence for phonology (the sound system), its records are too scanty to help much with the grammar (morphology or syntax), for which the best available evidence is Old Irish, the most archaic of the Insular languages. The records provide a picture of a language of the same type as Latin or Common Germanic; that is, one that still maintains a considerable part of the structure of the ancestral Indo-European language and has not lost final or medial syllables. Its vowel system differs only slightly from that reconstructed for Indo-European by the French linguist Antoine Meillet. Differences include the occurrence of Celtic **ī* for Indo-European **ē* (*e.g.*, Gaulish *rix* and Irish *rí*, "king"; compare Latin *rex*) and **ā* in place of **ō*. (An asterisk [*] before a letter or word indicates that the sound or word is not attested but is a hypothetical, reconstructed form.)

The consonantal system, too, is conservative, although there are some striking features. Among them are the loss of **p* (*e.g.*, Irish *athair* "father"; *cf.* Latin *pater*) and the falling together of the aspirated and unaspirated voiced stops assumed for Indo-European. (A stop is a consonant made with complete momentary stoppage of the breath stream some place in the vocal tract; voiced stops are those produced with the vocal cords vibrating, such as *b, d, g.* An aspirated sound is accompanied by a puff of breath, often written as an *h*, as in *bh, dh, gh;* an unaspirated consonant lacks this accompanying puff of breath.) Thus, Old Irish *dán* "what is given" corresponds to Latin *donum* "gift" (from Indo-European **d*), but Old Irish *de-naid* "sucks" corresponds to Latin *fe-* in *fe-mina, fe-llare* (from Indo-European **dh*). This loss of distinctive aspiration occurs with three out of the four voiced stops, a situation close to that of Slavic.

Other considerations, however, show that Celtic belongs to the so-called southern group of the European branch of Indo-European languages, or in another classification, to the same centum group as Latin, whereas Slavic belongs to the satem group. (The centum and satem divisions of Indo-European languages are made according to the treatment of certain sounds, called palatals, that existed in the ancestral Indo-European language.)

Correspondence of Celtic *p* and *q*

The loss of **p* in Celtic was very early; only the place-name Hercynia, preserved in Greek, shows that, in initial position, it became an *h* sound before disappearing. In most of the known Celtic languages, a new *p* sound has arisen as a reflex of the Indo-European **kʷ* sound. Thus there is Gaulish *pempe*, Welsh *pimp* "five," compared to Old Irish *cóic* and Latin *quinque* "five." The Irish evidence shows that **kʷenkʷe* must be reconstructed as the form in Common Celtic. The terms P-Celtic and Q-Celtic are sometimes used to describe assumed divisions of Common Celtic; to use one sound shift to distinguish dialects is, however, hardly justified, and the classification will not be used in this article.

The morphology (structure) of nouns and adjectives shows no striking changes from Indo-European. The Irish verb, however, exhibits a remarkable archaism not found in any other recorded Indo-European language. It has recently been demonstrated that the so-called primary and secondary endings of the Indo-European verb, as in the 3rd person singular endings **-(e)t* and *-(e)ti*, both occurred in the same tense. The forms with **-i* were used when the verb had absolute initial position; those without it were used in the normal verbal position at the end of the sentence. This is reflected in the Old Irish forms *beirith* (from **bereti*) "he bears" and *ní beir* (from **beret*) "he does not bear." It cannot be stated with certainty that Continental Celtic had preserved such forms. The Continental Celtic dialects show a few cases of sentences—admittedly imperfectly understood—in which the verb appears to be placed after the subject and before the object, as in modern western European languages. The history of Insular Celtic, however, shows a gradual shift from the older final position of the verb to the initial position, a position that has now become regular in all of the languages.

Relationships and ancient contacts of Celtic. The question of the relationship of Common Celtic to the other Indo-European languages remains open. For some time, it was held that Celtic stood in an especially close relation to the Italic branch; some scholars even spoke of a period when an Italo-Celtic "nation" existed, toward the end of the 2nd millennium BC. The existence of a *q–p* relationship (see above) inside Italic too (*e.g.*, Latin *quattuor* "four," but Oscan *petora*) was thought by some to support this view. Much of this argument is, however, based on accidental resemblances (*e.g.*, the Irish future tense in *f-* and the Latin future in *b-*) or on formations such as the deponent and passive verb forms ending in *-r*, which at one time were known mainly in Italic and Celtic but have since been found in the Hittite and Tocharian languages as well. The undeniable common features between Celtic and Italic, such as the superlative endings of adjectives (Latin *-issimus;* Celtic **-samos, *-isamos*), are hardly sufficient to justify the assumption of a special relationship, and the whole concept of an Italo-Celtic unity has been powerfully criticized by the linguists Carl J.S. Marstrander and Calvert Watkins.

The original home of the Celts cannot be located precisely, but, on the whole, the evidence points to the eastern part of central Europe. There is more evidence for their contacts with other Indo-European peoples. One group of Celts, at least, found themselves neighbours of the Germanic peoples and were often confused with them by classical writers. It can be inferred that the Celts had attained a higher standard of social organization than the Germanic peoples from the existence of words such as Gothic *reiki* and *andbahts* (modern German *Reich, Amt*), apparently borrowed from Celtic **rīgion* "kingdom" and **ambactos* "officer." To the Greeks and Romans, on the other hand, the Celts were inferior in culture; Celtic words in Greek are restricted to those describing Celtic institutions, such as *bardoi* "poets." The borrowings from Celtic into Latin, which derive mainly from the period before the expansion of Roman power, belong to a few restricted categories, such as war (*lancea* "lance"), transport (*carrus* "baggage wagon" and *carpentum* "light carriage"), and agricultural products (*cervesia* "beer"). When the Romans finally conquered Gaul and imposed their language, a number of Celtic words came into Latin, but the Celtic terms were mainly concerned with rural life. These are more common in French dialects than in standard French, which preserves a mere handful, such as *mouton* "sheep," *ruche* "beehive," and *arpent* "land measure."

Contacts with the Greeks and Romans

Early records of Celtic. *Continental Celtic.* Celtic died out very quickly in eastern Europe. In its farthest outpost in Asia Minor, it may be assumed that the *Letter of Paul to the Galatians* was addressed to a people whose culture was already Greek but whose Celtic origins are clear from names preserved by classical authors; *e.g.*, Drunemeton "the very sacred place," formed from two distinctively Celtic elements. When commenting on that Letter, St. Jerome (died AD 419/420) said that the Galatians still spoke a language almost the same as that of the people of Trier. As he had been in both places, his evidence cannot be dismissed offhand, and it may be that a few speakers of Celtic still existed in both areas. Since St. Jerome did not claim to know any Celtic dialect, however, his statement cannot be accepted with certainty. Speaking generally, the history of Continental Celtic comes to an end as that of Insular Celtic begins.

Insular Celtic. The earliest evidence for Insular Celtic consists, like that for Continental Celtic, mainly of names recorded by Greek and Latin authors. In the case of Ireland, these were entirely by hearsay, and many of the Irish place-names recorded by Ptolemy in the 2nd century AD have not yet been identified. From perhaps the 4th century, ogham inscriptions (see WRITING) are found in Ireland, consisting almost entirely of personal names. From the 5th century onward, British names in Latin inscriptions are recorded in Wales, as well as Irish names in both Latin and ogham alphabets in areas of Irish settlement. These scanty records are of value above all in establishing that, up to very nearly the time at which written documents become available, British and Irish had remained similar in structure to Gaulish. Thus, in Britain is found the genitive (possessive form) Catotigirni, which in Old Welsh gives Cattegirn, and in Ireland the genitive Dovatuci exists, which in Old Irish gives Dubthaich. These changes—loss of final syllables and connecting vowels, weakening of consonants between vowels, and so on—are very similar to what was happening to Latin in France at the same time; *e.g.*, Latin *avicellus* and *aqua* finally became *oiseau* and *eau*. There is no satisfactory explanation of why these profound changes should have occurred at this time, nor can the period at which they occurred be fixed precisely, for the engravers of the inscriptions clearly went on using traditional forms long after the sound had changed.

LINGUISTIC CHARACTERISTICS OF THE INSULAR CELTIC TONGUES

The new languages, the only forms of Celtic that are known thoroughly, present a considerable number of unusual features, some of them unknown to other Indo-European languages. Some scholars have argued that these features may have resulted from the presence of a large non-Celtic substratum in the British Isles. Because it is hardly likely that the Celtic invasions of those islands began much before 500 BC or that the invaders exterminated the existing inhabitants, such a possibility cannot be denied. On the other hand, some features once thought to be exotic, such as the initial position of the verb in the sentence, have been convincingly demonstrated to be organic developments from Indo-European. Others, such as the system of counting by 20s, are clearly innovations, but this system is shared by English ("three score and ten"), French (*quatre-vingts* "80"), and Danish, in all of which it is also an innovation, as well as Basque, in which it appears to be old.

Possible influences of substratum languages

Phonological characteristics. The most remarkable phonological feature of Insular Celtic is the development of a double series of consonants in which strongly articulated consonants are distinguished from their weak counterparts. The two series were originally merely phonetic variants, with the strong variety occurring in absolute initial position and in certain consonant clusters and the weak elsewhere. Later, however, the two series became independent, or phonological. In the languages as they first appear in writing, considerable changes have taken place in the phonetic forms of the two series. Both in Irish and in Welsh, Cornish, and Breton, the opposition (contrast) of strong:weak in the voiced stops has been replaced by stop:spirant (*e.g.*, *b:v*). (A spirant, such as *v*, *f*, *s*, is produced with local friction and without complete stoppage of the breath stream.) Irish has the same system for the unvoiced stops (*e.g.*, *t:th*), but Welsh, Cornish, and Breton have voicing in this instance (*e.g.*, voiceless *t*:voiced *d*). These changes by themselves are not very different from the weakening of consonants between vowels that occurs in other western European languages (compare Welsh *pader* "prayer," a loanword from Latin *pater* "father," with Spanish *padre* "father," deriving from Latin *patrem*), but, in Insular Celtic, they occurred not only inside the word but also inside the phrase, so that the initial consonant of a word preceded by another word ending in a vowel was weakened. When the final syllables were lost in the evolution to the modern languages, these variations remained, and a system of initial mutations (changes) was thus set up. If, for example, a Goidelic nominative form **sindos kattos koilos* "the thin cat" is reconstructed, this will give Old Irish *in catt coel* after the loss of final syllables, but the genitive **sindī kattī koilī* "of the thin cat" will give *in chaitt choíl* with changed initial consonants. The same sort of change occurred in one Italian dialect: in Tuscan, there occur *porta* "door," *la forta* "the door," *tre porte* "three doors," from Latin *porta*, *illa porta*, *tres portae*. In both cases, consonant weakening has spread from word to sentence; there is a common development, but it cannot be claimed that it is distinctively Celtic.

Grammatical characteristics. Another feature of Insular Celtic is its lack of the infinitive form of the verb found in most other Indo-European languages—*e.g.*, English "to do," "to call." The equivalent is the verbal noun, which is a noun closely linked to the verb, though not necessarily derived from the same stem. Being a noun, it can have a following noun in the genitive case, which, in the older languages at least, is subjective or objective according to whether the verb with which it is linked is intransitive or transitive. Thus, from the Old Irish sentence *téit in ben* "the woman goes," the verbal noun phrase *techt inna mná* "the coming of the woman" can be derived, whereas from *marbaid in mnaí* "he kills the woman" can be formed *marbad inna mná* (*lais*) "the killing of the woman (by him)." Among many other functions of the verbal noun is its use, when preceded by the appropriate preposition, with the substantive verb to provide a tense with continuous meaning. Thus, to *téit in ben* there is a parallel *a-tá in ben oc techt* "the woman is at going" (= "the woman is going"), and to *marbaid in mnaí* corresponds *a-tá oc marbad inna mná* "he is killing the woman." The close resemblance of this system to that of modern English, in which it is a comparatively recent development, has been variously explained as the working of a substratum or, more recently, in terms of areal (regional) development.

Verbal nouns

MODERN LANGUAGES OF THE FAMILY

The discussion of the individual languages that follows divides them into the two main groups, beginning with Irish, which is the oldest attested.

Irish. The history of Irish may be divided into four periods: that of the ogham inscriptions, probably AD 300–500; Old Irish, 600–900; Middle Irish, 900–1200; and Modern Irish, 1200 to the present. This division is necessarily arbitrary, and archaizing tendencies confuse the situation, especially during the period 1200–1600, when a highly standardized literary norm was dominant. After 1600, the modern dialects, among them Scottish Gaelic and Manx, begin to appear in writing.

The Latin alphabet was introduced into Ireland by British missionaries in the 5th century and soon began to be used for writing Irish. By the middle of the 6th century, the process of putting into literary form the rich oral tradition of the native learned class was certainly well advanced. The problems of interpreting the early writings are complicated by the fact that the orthography was based on that of Latin, but with a British pronunciation; *e.g.*, Latin *pater* was read as *pader*, the form of the loanword in Modern Welsh, and Old Irish *Pátric* was read as *Pádraig* (as it is spelled in Modern Irish). No new letters were evolved; the weak (less forceful) consonants were distinguished only in instances in which there were Latin spellings that could be utilized (*e.g.*, strong *ll*: weak *l*, strong *rr*: weak *r*, *nn:n*, *c:ch*, *t:th*) or with the help of the *punctum delens* (*s:ṡ*, *f:ḟ*), a dot that shows that the sound is not pronounced. As a result, many ambiguities remain: *ní beir* can mean either "he does not carry" or "he does not carry it," according to whether the *b-* is read as a *b* sound or a *v* sound. Nor was the Latin alphabet capable of dealing with the new system of consonant quality that appears in Irish alone among the Celtic languages. Thus, from the Celtic nominative singular and plural forms *bardos*, *bardī* developed Welsh *bardd*, plural *beirdd*, with a vowel alternation like that of English "mouse, mice." In Irish, the forms are *bard*, *baird*; the *-i-* of *baird* is purely graphic, serving to indicate that the following consonants are both palatalized. (Palatalized consonants are those in which the pronunciation is modified by raising the tongue toward the hard palate.) This palatalization had been purely phonetic as long as the *-ī* that caused it survived, but in Old Irish the palatalization became independent, so that each consonant of Common Celtic evolved into four distinct consonants (*i.e.*, phonemes); for example, from original Common Celtic *b* are derived a *b* sound and a palatalized *b* sound, and a *v* sound and a palatalized *v* sound.

Development of palatalized consonants

Apart from these phonetic developments, Old Irish is striking chiefly for the extraordinary proliferation of particles that appear before the verb and are used in forming compound verbs. For example, the Latin word *suffio* "I fumigate" is translated as *fo-timmdiriut*, composed of *fo* "under," *to* "to," *imb-* "around," *di* "from," and the stem *reth-* "run," with vowel and consonant changes appropriate to the 1st person singular present tense. Such forms, combined with a system of infixed accusative and dative pronouns (*i.e.*, pronouns inserted within a word) and syntactical accent shifts, produced a verbal system almost as complicated as that of Basque, though transparently Indo-European in origin. This system began to break down during the Old Irish period; the process was no doubt accelerated by the Viking raids that began at the end of the 8th century and that disrupted the monastic system, the guardian of the literary norm of Old Irish. Popular forms broke through in the Middle Irish period, though always mixed with archaizing forms; the backward-looking Irish scribes were never content to write down their own vernacular. During the 12th century, many ecclesiastical synods were held with the object of bringing the organization of the Irish Church more closely into line with that of western Europe, and the Anglo-Norman invasion took place in the latter part of the same century. It may have been these far-reaching changes that inspired the Irish literati to undertake a new standardization of their language. From the beginning of the 13th century, there is a rigidly fixed norm, often called Classical Modern Irish, which, for over four centuries, was used as the exclusive literary medium in Ireland and in Gaelic-speaking Scotland (there is no evidence for the Isle of Man).

The Scandinavians were first contained and then absorbed; they contributed a small number of loanwords to Irish, mainly in the field of navigation but also in that of urban life, for they were the first to establish towns in Ireland, though only on the coast. The Anglo-Normans were a more serious problem. After almost complete success in the early period, however, they became largely Gaelicized in custom and language outside the towns they had founded. They contributed a large number of loanwords to Irish in the fields of warfare, architecture, and administration, though many of these were comparatively short-lived. When English took over from Anglo-Norman as the language of administration and English colonies began to be planted in Ireland, English loanwords began to come into Irish. Few of these, however, were recognized in the literary language, and only from the evidence of the modern dialects has it become clear that they were quite numerous.

It was not until the beginning of the 17th century that the English power was finally consolidated in Ireland, first by military conquest and later by the planting of English-speaking colonists on a much larger scale than before. From this time onward, the decline of Irish began, with Irish becoming the language of an oppressed people. With no schools to teach the literary language nor any native nobility to support the literati who used it, the dialects appeared for the first time and began to be written in the paper manuscripts that constituted almost the only form of publishing available to those using Irish. By the beginning of the 19th century, it is probable that the population was almost equally divided between Irish speakers, mainly in the western half, and English speakers, mainly in the eastern half. The real imbalance lay in the fact that many of the Irish speakers were bilingual, whereas few of the English speakers were. The first census to record language use was taken in 1851, after the great famine that had struck the western areas with exceptional severity. By this time, the total number of Irish speakers was 1,524,286 (23 percent of the population), but only 319,602 spoke Irish exclusively. The decline of Irish has continued to the present day, in spite of a revival campaign initiated by the Gaelic League in 1893 and made part of official policy after the establishment of the Irish Free State in 1921.

The decline of Irish

Since then, Irish has been recognized as the first official language of the state; it is a compulsory subject in all of the schools and is a requirement for civil service and some other posts. There are probably more people able to read Irish—perhaps 300,000—than there ever were before. From 1945 onward, a standard written language has evolved, and there is a small but flourishing literary movement. Nearly all of the readers of Irish are English speakers by upbringing, however, and not many of them would claim that Irish had become their main language. In the western areas in which Irish was the traditional speech, there are now fewer than 50,000 people to whom it is a mother tongue, and all but a handful of these have a more or less adequate command of English.

Scottish Gaelic. Some aspects of the modern Scottish Gaelic dialects show that they preserve features lost in the language of Ireland during the Old Irish period; such archaism is characteristic of "colonial" languages. The innovations are, however, more striking than the archaisms. Most remarkable is the loss of the voicing feature (*i.e.*, the vibration of the vocal cords) in the stops. All of the stopped consonants are unvoiced, and the original voiceless stops have become strongly aspirated; for example, the equivalent of Irish *bog* "soft" is [pok], *p* being the voiceless counterpart of *b*, and that of *cat* "cat" is [k^{h}aht], the superscript ^{h} after *k* indicating the aspirated quality. (The brackets indicate that the symbols printed within them are phonetic rather than orthographic.)

Scottish Gaelic was planted on British soil, and the verbal system has been remolded on the lines of the British language, which originally had no future tense. As in Modern Welsh, the inherited present tense has largely future meaning, and present time is mainly expressed by the present-tense form of the substantive verb and the preposition *a(ig)* with the verbal noun. (In Insular Celtic, there are two verbs for "to be," a substantive verb with the meaning, roughly, "to exist," and a linking verb such as "is" in "John is a boy" or "sky is blue.") Thus, from Old Irish *téit in ben* "the woman goes" is derived Scottish Gaelic *théid a' bhean* "the woman will go," and from Old Irish *a-tá in ben oc techt* "the woman is going" results the Scottish Gaelic form *thà an bhean a' dol* "the woman goes" or "the woman is going."

It is only from the 17th century onward that the development of Scottish Gaelic can be studied, for, up until then, Classical Modern Irish was the literary norm. Indeed, the first book to be printed in Irish was a translation of the Calvinist *Book of Common Order*, published in Edinburgh in 1567, and the Scottish Reformers used the Irish Bible for some time, until it became clear that it was too foreign for the people to understand. A native Scottish standard emerged gradually during the 17th century, as poets ignorant of the Irish norm began to compose in their native dialects. It was not until the 18th century that the orthography became more or less fixed, and, until recent reforms in Ireland, the divergencies between the written languages were comparatively small. It is clear, however, that Scottish Gaelic must now be regarded as a separate language, though the differences between it and Irish are no greater than those between standard German and the Swiss dialects.

Scottish Gaelic as a separate language from Irish

Scottish Gaelic was confronted by northern dialects of English (Scots) from the very beginning; these rapidly penetrated into the east of the country, especially in the area centred on Edinburgh, the capital. The so-called Highland Line, marking the boundary between the two languages, has been steadily receding to the west since medieval times. By 1901, there were 230,806 speakers of the language, including 28,106 who spoke Scottish Gaelic exclusively; 106,466 persons, including nearly all of the monolingual Scottish Gaelic people, lived in the two counties of Inverness and Ross. The decline has continued steadily, and, even in those two counties, Gaelic is rapidly disappearing from the mainland, though it is holding its ground well in the Hebrides. Scottish Gaelic speakers in the early 1980s numbered about 90,700, which shows that the state of Scottish Gaelic survival is in many ways less serious than that of Irish. Because the majority of Gaelic speakers are Protestants who are accustomed to reading the Bible and using the vernacular in their religious services, literacy in Gaelic has been widespread. Furthermore, however low the census figures may be, they give an accurate picture of the number of those to whom Gaelic is a mother tongue because the number of English speakers who have acquired it is negligible. It must be admitted, however, that the recent literary revival finds its audience among the displaced Gaelic speakers of Edinburgh and Glasgow rather than in the Hebrides, where Gaelic is still confined to the home and English is the language of culture. In addition, there were about 500 Gaelic speakers in Nova Scotia in the early 1980s.

Manx. The history of the Isle of Man is imperfectly known. It was first inhabited by British speakers, then colonized from Ireland, and later became part of the Scandinavian Lordship of the Isles until 1266, when the King of Norway ceded both Man and the Hebrides to Scotland. From then on, it became involved in the wars between England and Scotland until 1346, when it passed finally to England. Though an Irish dialect survived as the speech of the majority of the people, these circumstances were not propitious for literary contacts with Ireland, and Manx was apparently not written until the Welsh bishop John Phillips translated the Anglican *Book of Common Prayer* in 1610, using an orthography based on that of English. This orthography makes Manx difficult to understand for readers of Irish and Scottish Gaelic, to whom it is of considerable interest because it represents a dialect entirely free of literary influences. The orthography soon became fixed, and a far-reaching series of later phonetic changes made the written form a highly inaccurate representation of the final stages of the language. Phonologically, it has more in common with the eastern dialects of Irish than with Scottish Gaelic, but its morphology and syntax are much more like those of Scottish Gaelic, probably because of the common British substratum. Its tense system is similar to that of Scottish Gaelic and Welsh, and its use of periphrastic verb forms (*i.e.*, longer forms with several elements) with the auxiliary meaning "to do" goes further than either of these, especially in its final stages.

Extinction of Manx in the early 20th century

In the beginning of the 18th century, English was still not understood by most of the people, but during the 19th century the decline of Manx was rapid, and the census of 1901 showed only 4,419 speakers of the language, all bilingual. Twenty years later, the language had ceased to be used as a normal means of communication, but, until recently, investigators have been able to find old people capable of giving useful information.

British languages. Britain was thoroughly romanized, and it is clear that the British language itself had been much affected by Latin; on the level of vocabulary, such an everyday word as Welsh *pysg* "fish," for example, derives from Latin *piscis.* The vowel system lost independent vowel quantity, the length of vowels becoming determined by the structure of the syllable, a situation that also occurred when the later Latin developed into Romance. Even after the collapse of Roman rule, Latin retained the same prestige among British Christians that it had in the rest of the Western Empire. The Irish monks introduced to the British speakers the custom of writing down the vernacular language at about the end of the 8th century; they adapted the clumsy Irish orthography for that purpose. At this period, the British dialects were very close to one another and can hardly be classed as separate languages, though they soon began to diverge. Like Old Irish, they had lost their final syllables and had undergone many other changes from the state shown by the inscriptions. Notably, the languages show only the merest traces of the declension of the noun, although the verb preserves a full inflectional system (that is, it has a full series of endings). It is clear that no future tense existed in early British, though the separate languages were later to fill this gap by various means.

Welsh. Welsh is the earliest and best attested of the British languages. Although the material is fragmentary until the 12th century, the course of the language can be traced from the end of the 8th century. The earliest evidence may represent the spoken language fairly accurately, but a poetic tradition was soon established, and by the 12th century there was a clear divergence between the archaizing verse and a modernizing prose. The latter was characterized by a predominance of periphrastic verbal-noun constructions at the expense of forms of the finite verb. By this time, too, the forms corresponding to other Celtic and Indo-European present-tense forms had largely acquired future meaning; *e.g.*, Welsh *nid â* "he will not go" (future) contrasts with Irish *ní aig* "he does not drive" (present). The gap thus left was filled, as in Scottish Gaelic and Manx, by a construction involving the substantive verb and the verbal noun; *e.g.*, *y mae'r wraig yn myned* "the woman goes" or "the woman is going" is composed of the verb *mae* "is" and the verbal noun *myned* "going."

By the 14th century, prose and verse styles became more similar, the prose being less colloquial and the verse less archaic. This marks the beginning of modern literary Welsh, which was finally fixed by the Bible translation of 1588. Modern literary Welsh developed at a time when Welsh national identity was beginning to be seriously threatened by the close relations with England that followed on the accession of the Welshman Henry Tudor (Henry VII) to the English throne in 1485. Welsh was being written less and less, and the spoken language was being penetrated by English words. In 1536, the Act of Union deprived Welsh of its official status. By the beginning of the 18th century, the position of the Welsh language had fallen very low, though it was still the vernacular of the vast majority of the people. It was saved by the Methodist revival of the 18th century, which established schools everywhere to teach the people how to read the Welsh Bible and which brought the Bible itself, together with Welsh religious books, into almost every home. The literary language rejected most of the English loanwords that had come into the popular speech, and, by the 19th century, a highly literate Wales was equipped with reading material of every kind in the Welsh language. Meanwhile, however, the popular speech diverged further from the fixed literary norm, which was never spoken except in the pulpit or on the platform. Modern Wales has a literary language that no mother speaks to her child and widely differing dialects that appear in print only to represent dialogue in stories and novels.

Effect of the 18th-century Methodist revival on Welsh

The Industrial Revolution of the 19th century first undermined the dominance of Welsh in Wales: English-speaking workers were brought into the mines and factories in such numbers that they could not be absorbed linguistically. By 1901 English speakers outnumbered Welsh speakers for the first time. Out of a population of 2,012,876, only 929,824 were reported as Welsh-speaking, though 280,985 people spoke Welsh alone. By the early 1980s the number of Welsh speakers had dropped to about 395,000, representing about 14 percent of an increased population. Most of rural Wales, however, is still Welsh speaking, and recent years have seen a great improvement in the official status of Welsh and a considerable increase in its use in the schools; it is certainly the most firmly rooted of the modern languages of Celtic origin.

In addition, there are still about 8,000 Welsh speakers in parts of Patagonia, Argentina, which was colonized by Welsh settlers in 1865. These people maintain cultural contacts with the homeland but are all bilingual in Welsh and Spanish and seem fated to final assimilation.

Breton. Breton disappeared from sight after the early period, and no literary texts are available until the 15th century. These, mainly mystery plays and similar religious material, are written in a standardized language that is by now completely differentiated from Welsh and, to a lesser degree, from Cornish. The divergence between Breton and Cornish is largely a matter of the English loanwords in Cornish and the French loanwords in Breton. The present tense was retained in its original function, whereas a future and conditional were formed from the present and past subjunctive, respectively. Later, the Breton dialects became written and showed considerable divergencies in this form. Not until the 1920s was an attempt at standardization made, and even then it was necessary to adopt two norms. One was called KLT, from the initials of the Breton names of the dioceses of Cornouaille, Léon, and Tréguier, the dialects of which agree with Welsh and Cornish in having the stress accent on the next to the last syllable. The other norm was the dialect of Vannes in the south, which has the stress accent on the final syllable and many other distinctive features, at least some of which can be explained by its close contacts with French. More recently, two norms have been evolved to cover all four

Breton dialect norms

dialects; one of these is used by most writers, whereas the other is officially recognized by the universities of Brest and Rennes, in both of which Breton is taught.

Up until recently, Breton was the common language of the people in Cornouaille, Léon, Tréguier and Vannes, within the boundaries of the *départements* of Côtes-du-Nord, Finistère, and Morbihan. Breton may still have more speakers than Welsh, but this is quite uncertain because no language statistics exist for France. There is, however, general agreement that very few children today are being brought up speaking Breton. This is at least partly the result of French official policy, which in effect excludes the language from primary and secondary schools, though the poor economic opportunities in Brittany also play a part. The literary movement is, therefore, confined to an intelligentsia of perhaps not much more than 10,000 people, many of whom live outside Brittany. The overwhelming mass of the remainder of Breton speakers are literate only in French, and chances for the survival of Breton seem very poor.

Cornish. Like Breton, Cornish had no literary texts before the 15th century. Those that exist are mainly mystery plays, some of which are almost literal translations from English. Cornish is much closer in structure to Breton than to Welsh, but it has also been heavily influenced by English. At the beginning of the 18th century, there were still a number of areas in which Cornish was spoken, but it died out as a means of communication by the end of the century. (D.Gr.)

Baltic languages

The Baltic languages form a branch of the Indo-European language family and are more closely related to Slavic, Germanic, and Indo-Iranian (in that order) than to the other branches of the family. They comprise modern Lithuanian and Latvian (Lettish), the languages of the Balts inhabiting the eastern coast of the Baltic Sea, as well as the now extinct Old Prussian language, Yotvingian (also spelled Yatvingian, Jotvingian, Jatvingian), Curonian (Kurish), Semigallian, and Selonian (Selian); the speakers of this group are here referred to as the B-Balts. There also existed languages and dialects of the Balts (D-Balts) who lived east of the above-mentioned groups in the areas of the upper reaches of the Dnepr River.

Languages of the group. Because its dialects are more archaic in their forms than those of the other living Indo-European languages, Lithuanian is of particular importance in the study of comparative Indo-European linguistics. The language had 2,760,000 speakers in Lithuania in the early 1980s and several thousand speakers in Belorussia and Poland, and until 1945 there were several thousand Lithuanians in East Prussia as well. More than 675,000 Lithuanians live abroad, mostly in the United States. Lithuanian is sharply divided into dialects whose differences are quite marked. The two major ones are Low (or Western) Lithuanian, with three subdialects, and High (or Eastern) Lithuanian, with four subdialects. The Low dialect is spoken by the Lowlanders, who live in the west and along the Baltic Sea; High Lithuanian is spoken by the Highlanders, who live in the eastern (and greater) part of Lithuania. Standard Lithuanian, formed at the end of the 19th and the beginning of the 20th century, is based on the southern subdialect of West High Lithuanian.

Lithuanian, most archaic living Indo-European language

The language most closely related to Lithuanian is Latvian, spoken by 1,344,000 speakers in Latvia in the early 1980s and about 156,000 abroad, mostly in the United States. Latvian is divided into dialects, the major ones being the Central dialect, Livonian (also called Tahmian, or West Latvian), and High (or East) Latvian. Standard Latvian, established at the end of the 19th and the beginning of the 20th century, is based on the Central dialect.

By the 16th century the Selonians, Semigallians, and Curonians (Kurs), who lived in areas of Latvia and Lithuania, had completely lost their national identities and were assimilated by the Latvians and the Lithuanians. They left no written records. Nor did the Yotvingians (or Suduvians), who lived in southwest Lithuania and farther to the south (in the territory of the present-day Poland). They became extinct around the 16th–17th century, being assimilated by the Lithuanians in the north and the Slavs in the south. Information on the extinct Baltic languages is extremely scarce (mostly place-names). Only Old Prussian, of all the extinct Baltic languages, left any written records, and they are quite poor. The Prussians lived in East Prussia (*i.e.*, between the lower reaches of the Vistula and Neman [Lithuanian Nemunas] rivers on the Baltic coast). They became extinct (*i.e.*, were assimilated by the Germans) at the beginning of the 18th century.

Extinction of several Baltic languages

Linguistically, the Yotvingians were very closely related to the Prussians. They made up one ethnic Baltic group, commonly called the Western Balts, as opposed to the so-called Eastern Balts—the Lithuanians, Latvians, Selonians, Semigallians, and Curonians. The traditional terms Western Balts and Eastern Balts are inaccurate when used for all of the Balts—*i.e.*, including the Balts for whose languages there are no records (the D-Balts). These Balts, who were assimilated by Slavs in the 7th–14th century, lived in the upper reaches of the Dnepr.

Historical survey. Proto-Baltic, the ancestral Baltic language from which the various known languages evolved, developed from the dialects of the northern area of Proto-Indo-European. These dialects also included the Slavic and Germanic protolanguages (and possibly also Tocharian). The quite close historic relationship of the Baltic, Slavic, and Germanic languages is shown by the fact that they alone of all the Indo-European languages have the sound *m* in the dative plural ending (*e.g.*, Lithuanian *vil̃ká-m-s* "wolf," Common Slavic **vilko-m-ŭ*, Gothic *wulf-am*). (An asterisk [*] indicates that the following sound or word is unattested and has been reconstructed as a hypothetical linguistic form.) This relationship is suggested not only by the morphology and word formation but also by the vocabulary—*e.g.*, Lithuanian *draũgas* (Latvian *dràugs*) "friend," Old Church Slavonic *drugŭ*, Gothic *driugan* "to fulfill military service"; Lithuanian *vãškas* (Latvian *vasks*) "wax," Russian *vosk*, Old High German *wahs*. Probably the earlier close contact of the Balts and the Slavs with the Germanic tribes broke off around the 2nd millennium BC, when the Balts moved from the south (not, however, losing contact with the Slavs) and settled a large area of the eastern coast of the Baltic Sea and the upper reaches of the Dnepr.

Features common to Baltic, Slavic, and Germanic

Relationship between Baltic and Slavic. Because contact between the Balts and Slavs from the time of Proto-Indo-European was never broken off, it is understandable that Baltic and Slavic should share more linguistic features than any of the other Indo-European languages. Thus, Indo-European **eu* passed to Baltic *jau* and Common Slavic **jau* (which became *ju*)—*e.g.*, Lithuanian *liáudis* "people," Latvian *ļàudis*, Old Church Slavonic *ljudije*. Tonal correspondences are found between Lithuanian and Serbo-Croatian (a Slavic language of Yugoslavia), and there are also similarities in stress; *e.g.*, Lithuanian *dūmai* "smoke" and Russian *dym* have the stress on the root, as do Lithuanian *rañką* "hand" (accusative singular) and Russian *rúku*, while both Lithuanian *rankà* "hand" (nominative singular) and Russian *ruká* are stressed on the second syllable.

Baltic and Slavic have specific morphological features in common. Among them, for example, is the genitive plural form. In Lithuanian, *mūsų* "of us" (= Latvian *mūsu*), evolved from the older form **nūsōn*, which comes from Baltic **nōsōn* and corresponds to the genitive plural form in Common Slavic, **nōsōn*, from which developed Old Church Slavonic *nasŭ* "of us." Baltic also shares some syntactic features with Slavic; *e.g.*, the genitive case is used in place of the accusative with verbs expressing negation (Lithuanian *jis* nieko *nežino* "he does not know anything," Latvian *viņš* nekā *nezin*, Russian *on* ničego *ne znajet*). There are also many lexical items common to Baltic and Slavic. More than 100 words are common in their form and meaning to Baltic and Slavic alone, among them Lithuanian *bėgu* "I run," Latvian *bēgu*, Old Church Slavonic *běgą*; Lithuanian *líepa* "linden tree," Latvian *liẽpa*, Old Prussian *lipe*, Old Church Slavonic *lipa*; Lithuanian *rãgas* "horn," Latvian *rags*, Old Prussian *ragis*, Old Church Slavonic *rogŭ*.

In addition to these features common to all the Baltic and Slavic languages, there are certain quite archaic features that Slavic shares with Lithuanian and Latvian but not with Old Prussian. The most striking example is the genitive singular ending in Lithuanian *vilk-o* = Latvian *vilk-a* "of a wolf," which comes from Baltic *-ō, historically paralleled by the genitive singular ending in Common Slavic **vĭlk-ā*. Old Prussian, however, has a different ending for the same inflection (*deiw-as* "of God"). In some instances the Slavic languages, differing from Lithuanian and Latvian, come closer to Old Prussian; *e.g.*, the Prussian possessive pronouns *mais* "my, mine," *twais* "your, yours," *swais* "one's own" are different from Lithuanian *mãnas, tãvas, sâvas* and from Latvian *mans, tavs, savs* but similar to Old Church Slavonic *mojĭ, tvojĭ, svojĭ*.

It is possible to conclude that there was close contact between the Baltic and Slavic protolanguages at the time when they began to develop as independent groups (*i.e.*, from about the 2nd millennium BC) and that the Proto-Slavic area might have been a part of peripheral Proto-Baltic, although a specific part. That is, Proto-Slavic at that time was in direct contact with both the corresponding dialects of the peripheral Proto-Baltic area (*e.g.*, with Proto-Prussian) and the corresponding dialects of the central Proto-Baltic area. All this shows that the Proto-Slavic area of that time (south of the Pripyat River) was much smaller than the Proto-Baltic area. Proto-Slavic began to develop as a separate linguistic entity in the 2nd millennium BC and was to remain quite unified for a long time to come. Proto-Baltic, however, besides developing into an independent linguistic unit in the 2nd millennium BC, also began gradually to split. Among other things, the size of the Proto-Baltic area had an influence on the development of Proto-Baltic in that it considerably reduced contact between its dialects (see also *Slavic languages*).

Development of the individual Baltic languages. By the middle of the 1st millennium BC, the Proto-Baltic area was already sharply split into dialects. From the middle of the 1st millennium AD, the Baltic language area began to shrink considerably; at that time the greater part of Baltic territory, the eastern part, began to be inhabited by Slavs migrating from the south. The Balts there were gradually assimilated by the Slavs; complete assimilation probably occurred around the 14th century. One of these Baltic tribes, the Galindians (Goljadĭ), is mentioned in a chronicle as late as the 12th century. The protolanguage of the so-called Eastern Balts split into Lithuanian and Latvian (Latgalian) around the 7th century. The other languages of the so-called Eastern Balts became separated probably at the same time. Selonian and Semigallian could have been transitional languages between Lithuanian and Latvian. Only Curonian, which some consider to be a transitional language between East and West Baltic, might have developed somewhat earlier. Moreover, the name of the Curonians occurs in historical sources earlier (AD 853: Latin Cori) than the names of the other tribes of the so-called Eastern Balts.

Split between Lithuanian and Latvian

Old Prussian. In historical sources the Prussians are called Aistians from the 1st century AD (by Tacitus) until the 9th century AD (by the Anglo-Saxon seafarer Wulfstan). They are first referred to by their own name (by a Bavarian geographer using the form Bruzi, "Prussians") in the 9th century AD. About 1230 the Teutonic Order began to plunder the lands of the Prussians and finally conquered them and the Yotvingians (Suduvians) in 1283. From that time the slow extinction of the two Baltic groups began, with the Germanization of the Prussians being completed at the beginning of the 18th century.

The earliest Old Prussian (and, for that matter, Baltic) written record is a German-Prussian vocabulary—the so-called Elbing vocabulary, compiled about 1300 and extant in a copy dated around 1400. This vocabulary, consisting of 802 Old Prussian words (and the same number of German words), was written in a South Prussian dialect (in Pomesania). Somewhat poorer than the Elbing vocabulary is the vocabulary compiled by Simon Grunau, consisting of 100 Old Prussian (and German) words, written between 1517 and 1526. The most important Old Prussian written records are the three catechisms of the 16th century based on the dialects of Sambia and translated from the German; the first two catechisms, which are very short and anonymous, date from 1545, and the third catechism, or *Enchiridion*, dates from 1561 and was translated by Abelis Vilis (Abel Will), a pastor of the church at Pobeten (Pabečiai; modern Romanovo). The language of all the Old Prussian catechisms is rather poor: the translations are excessively literal, and there are many errors in language and orthography. In spite of this, it is from these Old Prussian catechisms that scholars can learn most about the Old Prussian language.

Lithuanian. Lithuanians are first mentioned in historical sources in AD 1009. Old Russian (more precisely, an East Slavic language based mainly on Belorussian), Latin, and Polish were used in official matters in the Grand Duchy of Lithuania, which was established in the mid-13th century and lasted until the 18th century. Lithuanian writings begin to appear in the 16th century, first in East Prussia (home to many Lithuanians) and, somewhat later, in the Grand Duchy of Lithuania. In East Prussia, a quite uniform written Lithuanian language, based on the West High Lithuanian dialect, had already been established by the second half of the 17th century. In Lithuania, however, a uniform written Lithuanian came into use only at the end of the 19th and the beginning of the 20th century, when a standard Lithuanian language, based on the (Southern) West High Lithuanian dialect (spoken in both East Prussia and Lithuania), was established. Martynas Mažvydas (died 1563), who published the first Lithuanian book (a catechism) in Königsberg (Lithuanian Karaliaučius; modern Kaliningrad) in the year 1547, is purported to be the first person to use Lithuanian as a written language. Others, in particular Baltramiejus Vilentas, Jonas Bretkūnas, and the pastor-poet Kristijonas Donelaitis, also took part in the formation and standardization of a written Lithuanian language in the 16th–18th century in East Prussia. Great influence was exerted by the first grammars of Lithuanian, by Danielius Kleinas (1653 and 1654), and the works of Donelaitis (1714–80), the first Lithuanian writer to become well known. In the Grand Duchy of Lithuania the first to use Lithuanian as a written language is held to be Mikalojus Daukša (died 1613), who published a catechism in 1595 and a prayer book (*Postilė*) in 1599. Later writers who helped to standardize written Lithuanian include Konstantinas Sirvydas, who prepared the first dictionary of Lithuanian (1629), Jonas Jaknavičius (1598–1668), and Saliamonas Slavočinskis (17th century). The works of Daukša and Sirvydas in particular, based on the Middle and East High Lithuanian dialects, did much toward establishing the practice of drawing on the various dialects in the creation of a written Lithuanian. This tradition waned in the 18th century but was revived at the beginning of the 19th, with the formation of a standard Lithuanian. The practice became most apparent at the end of the 19th and the beginning of the 20th century, during the establishment of standard Lithuanian. The mixing and levelling of the Lithuanian dialects started at the beginning of the 20th century owing to the influence of a standard language, and it was especially intensified after the creation of the Lithuanian Soviet Socialist Republic in 1940. Both the Lithuanian S.S.R. and its successor, the Republic of Lithuania (from 1991), designated Lithuanian the nation's official language.

Establishment of uniform written Lithuanian

Latvian. The Latvian (Latgalian) people achieved a separate identity around the 16th century AD, when they completely assimilated the other Balts, as well as a greater part of the Livs (also called Livonians, Livians), who are of Finnic descent and live on Latvian territory. As a result of the conquering of Latvian territory by the German Knights of the Sword by 1290, close contact between all of the so-called Eastern Balts (the Latvians with the Lithuanians as well) was considerably weakened for a long period of time. The first Latvian book was the *Catechismus Catholicorum* of 1585. In 1638 the first Latvian(-German) dictionary, by Georgius Mancelius, appeared; the first grammar of the Latvian language, by Johann Georg Rehehausen, was published in 1644; and a Latvian translation of the Bible was published in 1685. The Latvian writings of the 16th–

18th century are translations of religious works, as are the Lithuanian. The language of these Latvian works, however, is somewhat poorer than that of the Lithuanian writings of the same period. The works of the Latvians Juris Alunāns (1832–64) and Atis Kronvalds (1837–75) exerted a great influence on the development of a standard Latvian language, based on the Central dialect, at the beginning of the 19th century. Standard Latvian was finally established at the end of the 19th and the beginning of the 20th century, and the levelling influence of this standard language on the Latvian dialects began at this time. Standard Latvian is the official language of Latvia.

Characteristics of the Baltic languages. All of the Baltic languages are inflected. Old Prussian is the most archaic of the recorded Baltic languages (although it also has innovations of its own), and it differs considerably from Lithuanian and Latvian.

Old Prussian. In contrast to Lithuanian and Latvian, Old Prussian retained the Baltic diphthong *ei*—Old Prussian *deiws* "God," Lithuanian *diẽvas,* Latvian *dìevs;* Old Prussian *deinan* "day" (accusative singular), Lithuanian *dienà,* Latvian *dìena.* In place of Lithuanian *š* and *ž* (from Indo-European **k̑, *g̑,* and **g̑h*), however, Old Prussian, like Latvian (as well as Curonian, Semigallian, and Selonian), has *s* and *z*—thus, Old Prussian *assis* "axle," Latvian *ass,* Lithuanian *ašìs;* Old Prussian *(po)sinnat* "to confess," Latvian *zināt,* Lithuanian *žinóti* "to know." The cluster $s+j$ (and $z+j$) in Old Prussian, as in Latvian, passed to *š* (and *ž*): Old Prussian *schan* (from **sjan*) "this" (accusative singular feminine), Latvian *šùo* "this," Lithuanian *šią.* In contrast to Lithuanian and Latvian, Old Prussian did not replace the clusters $t+j$ and $d+j$ with affricate sounds (begun with complete stoppage of the breath stream from the lungs and released with incomplete closure and friction): Old Prussian *median* "forest," Lithuanian *medžias,* Latvian *mežs.*

Word stress was free in Old Prussian, as it is in Lithuanian (in contrast to Latvian, in which the stress is predictable and falls on the first syllable). Old Prussian also made use of intonations (tones), the character of which is similar to that of the Latvian (*i.e.,* more archaic than that of Lithuanian intonations). The Proto-Baltic circumflex intonation corresponds to the falling tone in Old Prussian, while the acute intonation corresponds to the rising tone.

Old Prussian morphology

Old Prussian, moreover, had a substantive neuter gender, lost by Lithuanian and Latvian: Old Prussian *assaran* "lake," Lithuanian *ežeras,* Latvian *ezers;* Old Prussian *lunkan* "bast," Lithuanian *lùnkas,* Latvian *lūks.* It differs in morphology from Lithuanian and Latvian in more than one instance—*e.g.,* in the genitive singular ending, Old Prussian *deiw-as* "of God" (Lithuanian *diev-o* = Latvian *diev-a*) and, in the dative singular, Old Prussian *tebbei* "to you" (Lithuanian *tavi* = Latvian *tev*), among others. Old Prussian did not have the dual number, only the singular and plural. Nouns were declined according to seven types. There were five cases: nominative, genitive, dative, accusative, and vocative. All verbs had three separate forms in the plural, but not in the singular. The 3rd person was the same in both the singular and the plural. There were three tenses: present, preterite, and future.

In vocabulary Old Prussian is quite similar to Lithuanian and Latvian (closer to Lithuanian than Latvian). It should be emphasized, however, that Old Prussian differs from Lithuanian and Latvian in that it retained a greater number of archaisms than either.

Comparison of Lithuanian and Latvian. The differences between Lithuanian and Latvian can be summarized in very broad terms by saying that Lithuanian is far more archaic than Latvian and that modern written Lithuanian could in many instances serve as a "protolanguage" for it. For example, Lithuanian has quite faithfully preserved the old sound combinations *an, en, in, un* (the same is true of Old Prussian, Curonian, Selonian, and, possibly, Semigallian), while they have passed in every case to *uo, ie, ī, ū* in Latvian; thus, Lithuanian *rankà* (Old Prussian *rancko*) = Latvian *rùoka* "hand," Lithuanian *peñktas* (Old Prussian *penckts*) = Latvian *piekt(ai)s* "fifth," Lithuanian *pìnti* = Latvian *pīt* "to weave, to twine," and Lithuanian *jùngas* = Latvian *jūgs* "yoke." The diphthongs *ei* (as well as *ai*) and *au* in final position were monophthongized and later shortened in Latvian—*e.g.,* Lithuanian *ved-eĩ* (2nd person singular preterite) = Latvian **ved-ie,* which became *ved-i* "you led"; Lithuanian *med-aũs* = Latvian **med-uos,* which became *med-us* "of honey." Long vowels at the end of polysyllabic words have been shortened in Latvian, and short vowels have been dropped—*e.g.,* Latvian *sak-a* "says" (which derives from **-ā*) = Lithuanian *sãk-o,* Latvian *pel-e* "mouse" (from **-ē*) = Lithuanian *pel-ė̃,* Latvian *vìlk-u* "wolf" (from **-uo*) = Lithuanian *viĺk-ą,* Latvian *daikts* "thing" (from **-ăs*) = Lithuanian *dáiktas,* and Latvian *nakts* "night" (from **-ĭs*) = Lithuanian *naktìs.* Palatalized *k* and *g,* formed with the blade of the tongue closer to the hard palate than nonpalatalized *k* and *g,* were retained in Lithuanian (as in Old Prussian and Semigallian) but changed to *c* (pronounced like *ts*) and *dz* in Latvian (as in Selonian and Curonian): Lithuanian *ãkys* "eyes" (Old Prussian *ackis*) = Latvian *acis,* and Lithuanian *gérvė* "crane" (Old Prussian *gerwe*) = Latvian *dzeŕve.* The change of the old clusters $t+j$ and $d+j$ progressed further in Latvian. Most Lithuanian dialects have *č* (as *ch* as in "church") and *dž* (as *j* in "jam"), whereas Latvian has *š* (as *sh* in "shore") and *ž* (as *z* in "azure")—*e.g.,* Lithuanian *trẽčias* "third" = Latvian *trešs;* Lithuanian *bríedžai* "elks" = Latvian *brìeži.* Another difference between Lithuanian and Latvian is that, instead of Lithuanian *š* and *ž,* Latvian (like Selonian, Semigallian, Curonian, and Old Prussian) has *s* and *z* sounds—*e.g.,* Lithuanian *širdìs* "heart" = Latvian *sirds;* Lithuanian *žiemà* "winter" = Latvian *zìema.* Proto-Latvian (and Prussian) $s+j$ and $z+j$ have passed to *š* and *ž*: Latvian *šūt* "to sew" = Lithuanian *siū́ti;* Latvian *eža* "of a hedgehog" (from Latvian **ezjā*) = Lithuanian *ežio.* Lithuanian has retained the initial clusters *pj* and *bj,* which in Latvian (and similarly in Slavic) have passed to *pļ* and *bļ*—*e.g.,* Lithuanian *piáuti* "to cut" (*pi* is pronounced as *pj*) = Latvian *pļaũt;* Lithuanian *biaurùs* "hideous, nasty" = Latvian *bļaũrs.*

Stress and tone in Lithuanian and Latvian

Lithuanian has a free stress in contrast to Latvian fixed stress, which occurs on the first syllable. Latvian is more archaic than Lithuanian in the intonations inherited from Proto-Baltic: the Proto-Baltic circumflex intonation has preserved its falling character in Latvian (it became rising in Lithuanian), and the Proto-Baltic acute intonation retained its rising character (it is falling in Lithuanian), although in some cases (because of stress retraction) it has been changed to the broken intonation; *e.g.,* Latvian *pìrsts* "finger" = Lithuanian *pir̃štas* (falling in Latvian and rising in Lithuanian from the Proto-Baltic circumflex), Latvian *vãrna* "crow" = Lithuanian *várna* (the rising or extended intonation in Latvian and the falling intonation in Lithuanian from the Proto-Baltic acute intonation), Latvian *zâle* "grass" = Lithuanian *žolė̇* (the Latvian broken intonation from the Proto-Baltic acute intonation through stress retraction).

There are really no differences in the older morphological features between Lithuanian and Latvian if one does not take into account innovations such as the Latvian debitive verb form (*man ir jāmācās* "I must study" or "it is necessary for me to study") and the Lithuanian frequentative past (*jie eidavo* "they used to go"). Lithuanian and Latvian have two grammatical genders (masculine and feminine) and two numbers (singular and plural), while some Lithuanian dialects also have the dual number. Both Lithuanian and Latvian have seven cases—nominative, genitive, dative, accusative, instrumental, locative, vocative. Standard Lithuanian has five declensions of nouns with 12 inflectional types; Latvian has six declensions with eight inflectional types. Lithuanian adjectives have three declensions, Latvian adjectives have one. The comparison of adjectives in the two languages is different. Both Lithuanian and Latvian have indefinite adjectives (Lithuanian *mãžas,* masculine, *mažà,* feminine, "a small one" = Latvian *mazs, maza*) and definite adjectives (Lithuanian *mažàsis, mažóji* "the small one" = Latvian *mazais, mazā*) with their own specific inflection. The verb in Lithuanian and Latvian has three conjugations (genetically different). There are three persons, the third of which is the same (apparently from the time of Proto-Indo-European) in both the singular and the plural (as well as the dual); for example:

Lithuanian		Latvian
Singular		Singular
1. *kertù*	("I cut, I strike")	1. *certu*
2. *kertì*	("you cut, you strike")	2. *certi*
3. *ker̃ta*	("he cuts, he strikes")	3. *cert*
Plural		Plural
1. *ker̃tame*	("we cut, we strike")	1. *certam*
2. *ker̃tate*	("you cut, you strike")	2. *certat*
3. *ker̃ta*	("they cut, they strike")	3. *cert*

Verb forms

The verb in Lithuanian and Latvian has three tenses (present, preterite [or past], future)—*e.g.*, Lithuanian *kertù,* Latvian *certu* (present); Lithuanian *kirtaũ,* Latvian *cirtu* (preterite); Lithuanian *kir̃siu,* Latvian *cir̃šu* (future). In contrast to Latvian, Lithuanian also has a frequentative past tense—*e.g., kir̃sdavau* "I used to cut, strike." Lithuanian and Latvian have many compound tense forms, compounded from the forms of the verb *búti* "to be" and participles. There are several moods in both languages, although moods in Lithuanian differ from those in Latvian. The system of participles (active and passive) in Lithuanian and Latvian is quite similar, although complicated—*e.g.,* Lithuanian *kertą̃s,* Latvian *certuošs* (present active); Lithuanian *ker̃tamas,* Latvian *certams* (present passive). Lithuanian and Latvian word order is quite free, and, in general, the syntax of both languages is quite similar.

Words are formed in Lithuanian and Latvian basically by means of suffixes, prefixes, and compounding. The languages are very similar in their early vocabulary, and the differences that do occur tend to be more of a semantic nature—*e.g.,* Lithuanian *mõša* "husband's sister" = Latvian *māsa* "sister"; Lithuanian *žam̃bas* "corner, angle (acute)" = Latvian *zùobs* "tooth." Some older lexical differences do occur, however (*e.g.,* Lithuanian *kraũjas* "blood" = Latvian *asins*; Lithuanian *sūnùs* "son" = Latvian *dēls*). In the newer vocabulary, there are now many differences between Lithuanian and Latvian.

Loanwords in Baltic. The Baltic languages have loanwords from the Slavic languages (*e.g.,* Old Prussian *curtis* "hunting dog," Lithuanian *kùrtas,* Latvian *kur̃ts* come from Slavic [compare Polish *chart*]; Lithuanian *mùilas* "soap" [compare Russian *mylo*]; Latvian *suods* "punishment, penalty" [compare Russian *sud*]). There are in addition a few loanwords from the Germanic languages, such as Gothic (*e.g.,* Old Prussian *ylo* "awl," Lithuanian *ýla,* Latvian *īlens*) and possibly Scandinavian as well. There are many loanwords directly from German; these are notable especially in Old Prussian and Latvian and are clearly a consequence of German colonization in the 13th century.

Relation between the Baltic and Finnic languages

The Balts first came in close contact with their northern neighbours, the Baltic Finns, about 2000 BC. This contact left traces in both the Baltic and the Finnic languages. Baltic has very few early loanwords from Finnic, but Finnic has many early loans from Baltic. Latvian, with many loanwords from Livonian and Estonian (both Finnic languages), has been more influenced by Finnic than has any other recorded Baltic language.

Orthography. The Lithuanian alphabet is based on the Roman (Latin) alphabet. It has 33 letters, several employing diacritical marks (*ą, č, ę, ė, į, š, ų, ū, ž*), and is phonetic. In linguistic literature an acute accent (´) is used for falling tones and a tilde (~) for rising tones; the grave accent (`) is used for short, stressed vowels.

The Latvian alphabet has 33 letters, 11 with diacritical marks: *ā, č, ē, ģ, ī, ķ, ļ, ņ, š, ū, ž.* A macron (¯) over a vowel indicates that it is long. In linguistic literature the following accents are used for the Latvian intonations: grave (`) for falling, tilde (~) for extended or rising, and circumflex (^) for broken.

The Old Prussian orthography is almost wholly based on the German orthography of that time and is quite inconsistent. Furthermore, every Old Prussian written record has its own specific orthography. (V.J.M.)

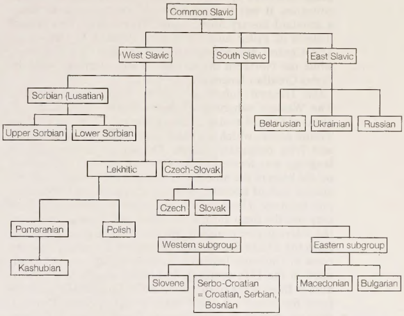
Slavic languages

The Slavic, or Slavonic, languages, a separate branch of the Indo-European language family, are closer to the Baltic languages than to any other Indo-European subgroup, but they share certain linguistic innovations with the other eastern Indo-European language groups (such as Indo-Iranian and Armenian) as well. From their homeland in east-central Europe (Poland or Ukraine), the Slavic languages have spread to the territory of the Balkans (Bulgarian, Macedonian, Slovene, Serbo-Croatian), central Europe (Czech and Slovak), eastern Europe (Belarusian, Ukrainian, Russian), and the northern parts of Asia (Russian). In addition, Russian is used as a second language by most inhabitants of the countries that were formerly part of the Soviet Union. Some of the Slavic languages have been used by writers of worldwide significance (*e.g.,* Russian, Polish, Czech), and the Church Slavonic language remains in use in the services in the Eastern Orthodox church.

LANGUAGES OF THE FAMILY

Three branches of Slavic

The Slavic language group is classified into three branches: the South Slavic branch, with two subgroups—Serbo-Croatian–Slovene and Bulgarian-Macedonian; the West Slavic branch, with three subgroups—Czech-Slovak, Sorbian, and Lekhitic (Polish and related tongues); and the East Slavic branch, comprising Russian, Ukrainian, and Belarusian.



In the spoken Slavic dialects (as opposed to the sharply differentiated literary languages) the linguistic frontiers are not always apparent. There are transitional dialects that connect the different languages, with the exception of the area where the South Slavs are separated from the other Slavs by the non-Slavic Romanians, Hungarians, and German-speaking Austrians. Even in this latter domain, some vestiges of the old dialectal continuity (between Slovene and Serbo-Croatian, on the one hand, and Czech and Slovak, on the other) can be traced; similar remnants of the old links are seen in comparing Bulgarian and Russian dialects.

Thus it should be noted that the traditional family tree of the Slavic group with three separate branches is not to be taken as the real model of historical development. It would be more realistic to represent the historical development as a process in which tendencies to differentiate and to reintegrate the dialects have been continuously at work, bringing about a remarkable degree of uniformity throughout the Slavic area.

Still it would be an exaggeration to suppose that communication between any two Slavs is possible without any linguistic complications. The myriad differences between the dialects and languages in phonetics, grammar, and above all vocabulary may cause misunderstandings even in the simplest of conversations; and the difficulties are

Linguistic complications

greater in the language of journalism, technical usage, and belles lettres, even in the case of closely connected languages. Thus, Russian *zelënyj* 'green' is recognizable to all Slavs, but *krasnyj* 'red' means 'beautiful' in the other languages. Serbo-Croatian *vrijedan* means 'hard-working,' but Russian *vrednyj* means 'harmful.' *Suknja* is 'skirt' in Serbo-Croatian, 'coat' in Slovene. The month *listopad* is October in Croatian, November in Polish and Czech.

South Slavic. *The Eastern subgroup: Bulgarian and Macedonian.* Bulgarian is spoken by more than nine million people in Bulgaria and adjacent areas of other Balkan countries and Ukraine. There are two major groups of Bulgarian dialects: Eastern Bulgarian, which became the basis of the literary language in the middle of the 19th century, and Western Bulgarian, which influenced the literary language. Bulgarian texts prepared before the 16th century were written mostly in an archaic language that preserved some features of both Old Bulgarian or Old Church Slavonic (10th to 11th century) and Middle Bulgarian (beginning in the 12th century).

Although the vocabulary and grammar of the early texts written in the Old Church Slavonic language include some Old Bulgarian features, the language was nevertheless based originally on a Macedonian dialect. Old Church Slavonic was the first Slavic language to be put down in written form. This was accomplished by SS. Cyril (Constantine) and Methodius, who translated the Bible into what later became known as Old Church Slavonic and invented a Slavic alphabet (Glagolitic). The modern Macedonian language is spoken by about two million people in the Balkan countries. It was the last major Slavic language to attain a standard literary form; during World War II its central dialects of Prilep and Veles were elevated to this status. The Central Macedonian dialect is closer to Bulgarian, while the Northern dialect shares some features with the Serbo-Croatian language.

The Western subgroup: Serbo-Croatian and Slovene. The Western subgroup of South Slavic includes the dialects of Serbo-Croatian, among them those of the Prizren-Timok group, which are close to some North Macedonian and West Bulgarian dialects. The literary Serbo-Croatian language was formed in the first half of the 19th century on the basis of the Shtokavian dialects that extend over the greater part of Bosnian, Serbian, Croatian, and Montenegrin territory. These dialects are called Shtokavian because they use the form *što* (pronounced in English as *shto*) for the interrogative pronoun 'what?'. They are distinguished from the Chakavian dialects of western Croatia, Istria, the coast of Dalmatia (where a literature in this dialect developed in the 15th century), and some islands in the Adriatic; in those areas *ča* (pronounced in English as *cha*) is the form for 'what?'. A third main group of Serbo-Croatian dialects, spoken in northwestern Croatia, uses *kaj* rather than *što* or *ča* and is therefore called Kajkavian. In all, some 20 million people use Serbo-Croatian (the present-day Croatian, Bosnian, or Serbian standard languages).

The Slovene language is spoken by more than 2.2 million persons in Slovenia and in the adjacent areas of Italy and Austria. It has some features in common with the Kajkavian dialects of Croatia and includes many dialects with great variations between them. In Slovene (particularly its Western and Northwestern dialects) some traces can be found of old links with the West Slavic languages (Czech and Slovak).

West Slavic. *Polish and other Lekhitic languages.* To the West Slavic branch belong Polish and other Lekhitic languages (Kashubian and its archaic variant Slovincian), Upper and Lower Sorbian (also called Lusatian or Wendish), Czech, and Slovak. More than 40 million people speak Polish not only in Poland and other parts of eastern Europe (notably in what are now Lithuania, the Czech Republic, and Belarus) but in France, the United States, and Canada as well.

Polish dialects

The main Polish dialects are Great Polish (spoken in the northwest), Little Polish (spoken in the southeast), Silesian, and Mazovian. The last dialect shares some features with Kashubian. The remaining speakers of Kashubian live west of Gdańsk near the Baltic Sea. Slovincian—now extinct—belonged to the Northern group of Kashubian dialects, which is distinguished from a Southern group. Kashubian dialects (including Slovincian) are considered to be the remnants of a Pomeranian subgroup within the Lekhitic group. Lekhitic also included Polabian, which was spoken up to the 17th–18th century by the Slavic population of the Elbe (Labe) River region. (At that time a dictionary and some phrases in the language were written down.)

Sorbian. The Polabian language bordered the Sorbian dialects, which are still spoken by inhabitants of Lusatia in eastern Germany. There are two literary languages: Upper Sorbian, used around Bautzen (Budyšin), and Lower Sorbian, used around Cottbus.

Czech-Slovak. Czech is spoken by some 12 million people in the Czech Republic; its dialects are divided into Bohemian, Moravian, and Silesian groups. The literary language is based on the 16th-century form of the Central Bohemian dialect of Prague. The Slovak literary language was formed on the basis of a Central Slovak dialect in the middle of the 19th century. Western Slovak dialects are similar to Moravian and differ from the Central and the Eastern dialects, which have features in common with Polish and Ukrainian. Some six million people speak Slovak; most live in Slovakia.

East Slavic: Russian, Ukrainian, and Belarusian. Russian, Ukrainian, and Belarusian constitute the East Slavic language group. Russian is spoken as a native language by some 160 million people, including many inhabitants of countries that were part of the former Soviet Union. Its main dialects are a Northern Great Russian group, a Southern Great Russian group, and a transitional Central group, including the dialect of Moscow, on which the literary language is based.

Ukrainian dialects are classified into Northern, Southeastern, Southwestern, and Carpathian groups (the last having features in common with Slovak); the literary language is based on the Kiev-Poltava dialect. More than 37 million people speak Ukrainian in Ukraine and neighbouring countries, and there are more than 350,000 Ukrainian speakers in Canada and the United States. Carpathian, also called Carpatho-Rusyn, has sometimes been considered a language apart. In 1995 a codified form of it (Rusyn) was presented in Slovakia, thus enabling the teaching of Rusyn in schools.

Some seven million people speak Belarusian in Belarus. Its main dialectal groups are Southwestern Belarusian, some features of which may be explained by contact with Polish, and Northeastern Belarusian. The dialect of Minsk, which served as a basis for the literary language, lies near the border between these two groups.

HISTORICAL SURVEY

Proto-Balto-Slavic. *Innovations.* Each branch of Slavic originally developed from Proto-Slavic, the ancestral parent language of the group, which in turn developed from an earlier language that was also the antecedent of the Proto-Baltic language. Both Slavic and Baltic share with the eastern Indo-European languages (called satem languages) the same change of Indo-European palatal *ḱ* and *ǵ* sounds (consonants produced by bringing the blade, or front, of the tongue up to or toward the hard palate, as in English *cue, argue*) into spirants of the *s* and *z* type (for example, in Proto-Slavic **sŭto* 'hundred' has an *s* sound contrasting with the *k* sound in Latin *centum*). (An asterisk indicates a reconstructed rather than an attested form.) The Slavic and Baltic branches are characterized by several innovations, including the change of the old Indo-European syllabic *r* and *l* (which functioned as vowels) to *ir* or *ur, il* or *ul;* and similar patterns of stress in nouns and verbs.

Innovations in Slavic and Baltic

Hypothetical origins. Some scholars believe that, after the common Indo-European area had divided into different dialect zones (after approximately 3000 BC), a protodialect developed in the Baltic and Slavic areas that had many features peculiar to only these two branches of Indo-European. At the same time this protodialect was connected with certain western Indo-European protodialects called Old European that are identified as the source of a number of river names. The ancient Baltic and Slavic names of rivers (hydronyms), such as the Russian Oka, are of the same type as the hydronyms found in central Europe.

The dialects of the Slavic protolanguage spoken near the Carpathian Mountains in the upper Vistula River area may have been part of the intermediate zone situated between the western Indo-European dialects (Germanic, Celtic, Italic, and so on) and the eastern Indo-European ones; in addition to Baltic and Slavic in the north, this intermediate zone included the ancient Indo-European languages of the Balkans (Illyrian, Thracian, Phrygian). The domain of the Proto-Balto-Slavic dialect may have been situated to the east of the Germanic and other Old European dialects, to the north of Ancient Balkanic, and to the west of Tocharian.

Domain of Proto-Balto-Slavic

The exact geographic borders of the Balto-Slavic domain appear impossible to determine, but they may well have been located in eastern Europe around present-day Lithuania and to the east and south of it. The later diffusion of Slavic languages southward into the Carpathian region may represent the spread of one of the dialects of this Old Baltic domain. The oldest Slavic protolanguage could be described as the result of further changes acting on the Baltic protolanguage (but not vice versa).

Until the middle of the 1st millennium AD, the Slavs were known to other peoples as the inhabitants of the vast territories between the Dnieper and Vistula rivers. In the 6th century AD the Slavs expanded to the Elbe River and the Adriatic Sea and across the Danube River to the Peloponnese (southern Greece). In that period, according to the oldest Greek and Latin writings about the Slavs, they were already divided into several groups. The Slavic language, however, was uniform in its phonological and grammatical structure, with important dialectal variations occurring only in the vocabulary. The main phonological difference between the oldest pattern common to Baltic and Slavic and the later one that characterized Slavic alone was that in Slavic all syllables became open (*i.e.,* a syllable could end only in a vowel). Thus, all consonants

Figure 16: Distribution of the Slavic languages in Europe.

at the end of a syllable were lost. This led to a reshuffling of most of the inflectional endings.

Term for 'king' as marker

An important clue to the date of the dissolution of Slavic unity is the separate development in different Slavic dialects of the name of the emperor Charlemagne (742–814). This name must have entered into Slavic in the postulated form *korljĭ 'Karl' before the dissolution took place. Subsequently the proper name became the generic term for 'king.' The segment *-or-* in the postulated form appears differently in the modern Slavic languages—compare Bulgarian *kral,* Serbo-Croatian *kralj,* Slovene *krâlj* (*i.e.,* South Slavic *-ra-*), Russian *korol'* (*i.e.,* East Slavic *-oro-*), Czech *král,* Polish *król.*

The loss of reduced vowels. The next period in Slavic linguistic history began with the loss of the "reduced" vowels *ŭ* and *ĭ,* called *yers,* that resulted from Indo-European short *u* and *i;* this loss caused a wide-ranging change in many words and forms. Although this process was common to all the Slavic dialects, which were still connected with each other at that period, it took place slowly and at different rates in different dialects, beginning in the 10th to the 12th century and expanding from the southwest to the northeast. With the loss of the *yers,* which gave different results in different dialectal groups (see Table 31), the uniformity of the Slavic language area finally disappeared, and separate branches and languages emerged.

The early development of the Slavic languages. The separate development of South Slavic was caused by a break in the links between the Balkan and the West Slavic groups that resulted from the settling of the Magyars in Hungary during the 10th century and from the Germanization of the Slavic regions of Bavaria and Austria. Some features common to Slovak and Slovene may have developed before the West-South break. The eastward expansion of dialects of Balkan Romanian (a Romance language) led to a break in the connection between the South and the East Slavic groups about the 11th–12th centuries. The history of the Balkan Slavs was closely connected with Byzantium, in contrast to that of the Lekhitic and Sorbian subgroups of the Western Slavs, which was connected with western European culture.

An effort on the part of the Slavs to counteract the influence of the Western Christian church (which was associated with the German empire) was the motive behind the introduction of the Old Church Slavonic language into the liturgy in Great Moravia, the first Slavic national state. Founded in the 9th century, Great Moravia united different groups speaking West Slavic dialects. In 863 its prince, Rostislav, invited St. Cyril and his brother St. Methodius to create a national church with a language and writing of its own. Prior to that time some Christian texts in Moravia might have been translated into Slavic from Latin (and partly perhaps from Old High German); these have been preserved only in later copies.

The disciples of Cyril and Methodius were soon forced to leave Moravia, and mostly they went south. The second period in the history of the Old Church Slavonic language (893–1081) occurred in the Bulgarian kingdoms of Symeon (893–927) and Peter (927–969) and in the kingdom of Samuel (997–1014), and it was connected with the literary activity of many Bulgarian scholars who translated numerous Greek texts into Slavic and also produced a small number of original works. In the writings of the period of Symeon and Peter, Western (Macedonian) features were replaced by Eastern (Bulgarian) ones.

Reflection of Old Church Slavonic variants in various manuscripts

Both the Western and Eastern variants (recensions) of the Old Church Slavonic language are preserved in manuscripts of the 11th century, while the East Slavic (Russian) variant is reflected in the oldest dated Slavic manuscript, *The Ostromir Gospel* (1056–57), and in many later texts. The Moravian variant must be reconstructed on the basis of some later texts (such as the Kiev fragments from the beginning of the 11th century), which were written after the break with the Great Moravian tradition.

In some documents of the 10th and 11th centuries, the Bohemian variant (which shares some West Slavic peculiarities with Moravian) has been preserved. Several features are common to the Moravian and Bohemian varieties of the Old Church Slavonic language, to the Slovene (Pannonian) variant reflected in the Freising fragments (late 10th century), and to the Croatian Old Church Slavonic tradition that is attested from the 12th century, as well as to the Serbian tradition. All these variants of Old Church Slavonic have some peculiarities that are to be explained as the result of the interaction of the original system with that of a local dialect. By approximately AD 1000 all Slavic languages were so similar to one another that such interaction was possible.

From these local variants of Old Church Slavonic that are preserved in the manuscripts of the 10th–12th centuries, one should distinguish the later local Church Slavonic languages (Russian, with its variants; Middle Bulgarian; Serbian, which in the 18th century was replaced in Serbia by the Russian variant; Croatian; and the Romanian variant of Church Slavonic, which was used as a literary and church language in Romania from the 14th to the 18th century). From the linguistic point of view, these later Church Slavonic literary languages differ from the earlier varieties chiefly in their systems of vowels; the early nasalized vowels were replaced by different later reflexes, and the reduced vowels (*yers*), with the exception of those followed by a syllable containing another *yer,* were generally lost. These changes in the sound pattern were accompanied by a number of culturally determined changes in vocabulary.

The emergence of the individual Slavic languages. After the schism between the Eastern (Orthodox) and Western (Roman) Christian churches in the 11th century and the beginning of the Crusades, the Church Slavonic language fell out of use in all West Slavic countries and in the western part of the Balkan Slavic region. The only exception was the renaissance of Croatian Church Slavonic in the 13th century. At the end of the same century, the first Czech verses in the local dialect were written; they were the precursors of the rich poetic literature in the Old Czech language that appeared in the 14th century. The early Czech literary language was marked by the influence of Latin, which had replaced the Bohemian variety of Old Church Slavonic as a literary language.

In the earliest period of its development, the Polish literary language was modeled on the Czech pattern. After the Christianization of Poland, Latin (and later German) loanwords entered the Polish language in their Czech form. The Czech influence is seen in the Polish literary language until the 16th century (the "Golden Age"), when Renaissance tendencies resulted in the creation of genuinely literary works more closely reflecting everyday speech. Later, the Polish literary language was enriched by cross-fertilization with Ukrainian and Belarusian.

The decline of western variants of Church Slavonic

In the 16th century in Dalmatia, poets who were influenced by the Italian Renaissance (and who also wrote in Italian and Latin) created a rich poetic literature in Croatian. A Slovene translation of the Bible was published in 1575–84, and Kashubian and Sorbian religious texts were also produced during this period. The comparatively early rise of the West Slavic (and the westernmost South Slavic) languages as separate literary vehicles was related to a variety of religious and political factors that resulted in the decline of the western variants of the Church Slavonic language.

In contrast, the continuing use of Bulgarian Church Slavonic and different variants of Russian Church Slavonic made it difficult to construct literary languages for Bulgarian and Russian that were based on everyday speech. Bulgarian texts were written in Bulgarian Church Slavonic until the 16th century. After that the so-called Damaskin religious literature developed, closer to the popular speech; its development, however, was hampered under Turkish rule. Most of the Old East Slavic (Old Russian) literary texts were written in a mixture of Russian Church Slavonic and the Old Russian vernacular language; only a few documents, particularly some parts of the chronicles (annals), are written entirely in Old Russian. The proportion of South Slavic (Church Slavonic) and East Slavic (Old Russian) elements in each text is different depending on its stylistic peculiarities.

In the middle of the 17th century, the old Great Russian variant of the Church Slavonic language in the official

Orthodox Church was replaced by a new variant taken from the southwestern East Slavic tradition, a form that incorporated some Ukrainian and Belarusian elements. This development was connected with a split in the Russian Orthodox church; the Old Believers, who split off from the main body of the church, continued to use the archaic Great Russian variant, while Patriarch Nikon's new variant, based on the southwestern tradition, was adopted by the official church and is used in it to this day. Because the Ukrainian tradition includes many West Slavic elements, this reform, which occurred after the incorporation of Ukraine into the Russian Empire, was a step in the direction of the Westernization of the Russian language that took place about 1700, when Tsar Peter I the Great began his attempts to reconstruct and Westernize the whole Russian way of life.

In the 18th and 19th centuries, many waves of loanwords from different Western languages entered the Russian language. During an earlier period Russian sentence structures had been formed on Germanic and Latin patterns; the intensive French-Russian bilingualism of the Russian elite in the 18th and 19th centuries not only influenced syntax but also brought a shift in the range of meanings of Russian words as the elite came into contact with western European concepts. The great Russian literature of the 19th and early 20th centuries (from Aleksandr Pushkin [1799–1837] up until Leo Tolstoy's death in 1910) created a literary language close to everyday speech, especially to that of the villages. In the official style of Russian, however, Church Slavonic elements are still widespread, as can be seen even in general newspaper articles.

A Pan-Slavic language

The concept of a language that would unite all the Slavs has remained in the back of the Slavic consciousness, not as a real aim but rather as an important symbol. An early interesting attempt to unite different chronological and local Slavic strata was carried out by the 17th-century Croatian traveler Juraj Križanić. In modern literature one might cite the experiments at unification of Velimir Khlebnikov, a Russian Futurist poet, and of the Polish poet Julian Tuwim, who invented words based on Russian and other Slavic roots in some of his poems.

The modern Slavic languages. Among the Slavic languages that attained their standard literary form at a later stage in Slavic history than those mentioned above is Ukrainian. It was used in some literary texts in the late 18th century and in turn influenced the language of Nikolay Gogol, one of the greatest Russian writers of the 19th century. In the 19th century and especially in the first decades of the 20th century, a number of great poets wrote in Ukrainian, notably Taras Shevchenko (1814–61) and Lesya Ukrainka (1871–1913). The movement toward national liberation led to the introduction of many neologisms into the language, which persisted even after the advent of Russian pressure to bring the languages closer again. After World War I, the Belarusian language became a standard language in the Belorussian Soviet Socialist Republic (now Belarus).

Since the breakup of the Soviet Union, Czechoslovakia, and Yugoslavia, all the Slavic languages have acquired the status of the main language of an independent state. Only the minor languages are exceptions: *e.g.,* Kashubian is used officially only in some cultural performances, and Upper and Lower Sorbian are taught in local schools in eastern Germany. The extent of dialectal variation in the different languages ranges from a very great degree in Slovene to a much smaller degree in Polish and Russian. Radio and other mass media have been among the main influences leading to linguistic consolidation. Languages such as Polish, Czech, and Russian, which have served as a basis for great literatures, have become models for others that are only now being put to literary use (although for such languages as Kashubian and, to some degree, for Sorbian, the folk literature remains much more important as a model than individual literary works and translations of past centuries).

LINGUISTIC CHARACTERISTICS

Common features. A number of features set off Slavic from other Indo-European subgroups. The Slavic languages are an unusually numerous yet close-knit subgroup. On the whole, Slavic auxiliary words tend to be unstressed and to be incorporated into a single phonetic group or phrase with an autonomous stressed word. Inflection (*i.e.,* the use of endings, prefixes, and vowel alternations) has persisted as the main method of differentiating grammatical meanings, although to a lesser degree in nouns than in verbs because many functions of the noun case endings may also be expressed by prepositions. Endings are largely fusional (*e.g., -te* means simultaneously 'second person' and 'plural'). Slavic more than other languages shows verb aspect overtly. The movable stress pattern common to most South and East Slavic languages has profoundly influenced their versification.

Many linguistic devices found both in the oral tradition and in literary works of the different Slavic languages may be traced to common ancestral forms. An exuberant use of diminutives and metaphoric figures marks the Slavic oral tradition. It seems possible to reconstruct a common Proto-Slavic model of the universe as seen through language. The main feature of such a model is recurring binary (two-way) contrasts, as is evidenced by such key words as *bogŭ* 'god' from 'a portion allotted by the gods' and *ne-bogŭ* 'not having its portion, having bad fortune.' Such pairing of opposites recalls the ancient Iranian dualistic view of the world, a view that evidently influenced the Slavs to a degree not yet fully appreciated.

Influence of other cultures on Slavic vocabularies

As compared with the common Indo-European scheme, the pre-Slavic cultural vocabulary seems somewhat simpler, evidently as a result of the loss of direct contact with the Southern civilizations that served as a pattern for pre-Indo-European culture. Later developments were caused largely by western European and Greek (particularly Byzantine Christian) influences and by contact with Asian and Middle Eastern cultures, which led to innovations in the vocabularies of the East Slavic and South Slavic languages. In some instances, whole series of terms for objects were borrowed into Russian and other East Slavic languages from eastern sources.

All Slavic languages are synthetic, expressing grammatical meaning through the use of affixes (suffixes and, in verbal forms, also prefixes), vowel alternations partly inherited from Indo-European, and consonant alternations resulting from linguistic processes peculiar to Slavic alone. Although analytical methods of expressing grammatical meanings (through prepositions and other "empty" grammatical words) are present in older strata of the language, they are used to the exclusion of all other means only in the case system of Modern Bulgarian and Macedonian. The tendency toward analytic expression is noticeable in everyday Russian speech, but the drift of the Slavic languages in this direction (as in the development of the western European languages) has been held back by the stabilization of the language resulting from mass communication and education.

Phonological characteristics. *Palatalization.* The systems of sounds in Slavic languages are rich in consonants, particularly in spirants (fricatives, like English *s, z, sh*) and affricates. This is especially true in comparison with the protolanguage and with other Indo-European languages. The affricates (which are consonant sounds like English *ch, ts,* begun as stops, with complete stoppage of the breath stream, and released as fricatives, with incomplete stoppage) have resulted historically from a succession of different processes of palatalization that have occurred in Slavic and are one of the most characteristic features of Slavic phonology.

Palatalization is the process whereby the pronunciation of an originally nonpalatal sound is changed to a palatal sound by touching the hard palate with the tongue; it is also the process whereby a nonpalatal sound is modified by simultaneously moving the tongue up to or toward the hard palate. Originally, palatalization was connected with the adaptation of a consonant to the following vowel within a syllable, specifically, with the adaptation of a consonant to a following front vowel. This adaptation gave rise to "soft" (palatalized) syllables, composed of palatalized consonants followed by front vowels. The *j* sound, as *y* in English *year* (from older nonsyllabic Indo-European

Table 29: Development of Proto-Slavic *tj, *dj

Proto-Slavic	Old Church Slavonic	South Slavic				East Slavic	West Slavic		
		Bulgarian	Macedonian	Serbo-Croatian	Slovene	Russian	Polish	Czech	Upper Sorbian
*tj, later *t'	št	št	ḱ	ć	č	č	c	c	c
*dj, later *d'	žd	žd	ǵ	đ	j	ž	dz	z	z
*světja 'candle'	svěšta	svešt	sveḱa	svijeća	sveča	sveča	świeca	svíce	swěca
*medja 'bound(ary)'	mežda	mežda	meǵa	međa	meja	meža	miedza	meze	mjeza

An asterisk (*) indicates an unattested, reconstructed form.

Development of palatals, affricates, and spirants

i), tended to palatalize the preceding consonant either by merging with it or by giving rise to consonant groups such as *bl'* from *bj* (*by*). As palatalized stop consonants (for instance *k', g', t', d'*) became increasingly differentiated from the corresponding nonpalatalized series (*k, g, t, d*), the palatalized stops tended to develop further into affricates (with the subsequent development of voiced affricates into spirants). Thus, palatalized *k'* before the ancient front vowels developed into the affricate *č* (as *ch* in English *church*), and palatalized *g'* in the same environment changed to *dž* (as *j* in *judge*), which became the spirant sound *ž* (as *z* in *azure*) in all Slavic languages.

Before front vowels resulting from ancient diphthongs, palatalized *k'* changed to a *ts* sound, written as *c* (*e.g.,* Old Church Slavonic *cěna* 'price,' Serbo-Croatian *cijèna,* Russian *cena,* cognate to Lithuanian *káina*), and *g'* changed to a *dz* sound, which later changed to *z* (Old Church Slavonic [*d*]*zelo* 'very,' Old Czech *zielo,* Belarusian dialect *do zěla,* cognate to Lithuanian *gaila*). The sounds *t'* (from *tj*) and *d'* (from *dj*) changed into different stops, affricates, and spirants in the separate Slavic languages (see Table 29).

These processes of assibilation of the palatalized velar (*k', g'*) and dental (*t', d'*) sounds happened repeatedly in the history of the individual Slavic languages. Palatalization (softness) as a distinctive feature of most consonant sounds has been preserved in East Slavic; for example, in Modern Russian palatalized (or soft) *t', d', s', z'* contrast with nonpalatalized (or hard) *t, d, s, z.* (The contrast between the palatalized *k'* and the hard *k* is just now in the process of development.) Some West Slavic languages also have this contrast of palatalized and nonpalatalized consonants, while others do not. Czech, Slovak, and Serbo-Croatian, which have the usual three sets of labial, dental, and velar consonants inherited from the protolanguage, have developed a special, additional series of palatal stops. In all the Slavic languages, voiced stop and fricative consonants (pronounced with vibrating vocal cords) contrast with voiceless consonants (pronounced without vibrating vocal cords).

The tendency to increase the number of different spirants (nonstops) is connected with the processes of palatalization. In Ukrainian and the Southern Russian dialects and in Belarusian, Czech, Slovak, Upper Sorbian, and some Slovene dialects there also developed a voiced velar spirant sound, forming a pair with the voiceless velar spirant *x* of the Proto-Slavic language. The nasal vowels *ę* and *ǫ* that had developed in Proto-Slavic from older combinations of vowels with nasal consonants (still retained in Baltic) have been preserved only in some Lekhitic languages and in some South Slavic dialects, especially those of Slovene (see Table 30).

Slavic vowel systems

The vowel systems are especially rich in those Slavic languages that have preserved prosodic differences in pitch (tone) and quantity (length versus shortness)—Serbo-Croatian, Slovene, and Northern Kashubian. The reshaping of the Slavic vowel systems is in large measure a result of the loss of the *yers,* which had different effects in different dialects (see Table 31).

Table 31: Results of Loss of *Yers*

English translation	Proto-Slavic	Russian	Bulgarian	Macedonian	Serbo-Croatian	Czech	Polish	Upper Sorbian	Lower Sorbian
'day'	*dĭnĭ	den' (d'en')	den	den	dan	den	dzień	dzeń	źeń
'dream'	*sŭnŭ	son	sŭn	son	san	sen	sen	son	son

An asterisk (*) indicates an unattested, reconstructed form.

Stress accents. Differences in vowel quantity have also been preserved in Czech and Slovak, in which new long vowels developed as a result of contraction. A fixed stress accent is found in the West Slavic languages as well as Macedonian, in contrast to Proto-Slavic, Serbo-Croatian, Slovene, Bulgarian, and the East Slavic languages. In Czech and Slovak, as well as in Sorbian and Southern Kashubian, stress is fixed on the first syllable of the word, but in Polish, Eastern Slovak, and Southern Macedonian, it falls on the next to the last syllable of the word, while in Western and standard Macedonian it falls on the third syllable from the end. The Slavic languages with a nonfixed placement of stress reflect the Proto-Slavic (and Indo-European) distinction between two types of noun and verb paradigms: (1) the paradigm with movable stress in which the stress (indicated here by ´) falls on the root in some forms and on the inflectional ending in others (*e.g.,* 'head' in Russian is *golová* in the nominative case and *gólovu* in the accusative; these forms derive from Proto-Slavic **golvá, *gólvǫ*) and (2) the paradigm in which the stress is fixed on the stem (*e.g.,* 'willow' in Russian is *íva* in the nominative case, *ívu* in the accusative, from **íva, *ívǫ*).

Table 30: Development of Proto-Slavic Nasal Vowels Compared with Baltic

English translation	Lithuanian	Proto-Slavic	Polish	Russian	Bulgarian
'five'	penki	*pętĭ	pięć	p'at'	pet
'hand, arm'	ranka	*rǫka	ręka	ruka	rŭka

An asterisk (*) indicates an unattested, reconstructed form.

Grammatical characteristics. *Cases.* Most Slavic languages reflect the old Proto-Slavic pattern of seven case forms (nominative, genitive, dative, accusative, locative, instrumental, vocative), which occurred in both the singular and the plural. There was also a dual number, meaning two persons or things. In the dual, the cases that were semantically close to each other were represented by a single form (nominative-accusative-vocative, instrumental-dative, genitive-locative). The dual is preserved today only in the westernmost area (*i.e.,* in Slovene and Sorbian). The trend toward the modern, more analytical type of construction using prepositions and away from the synthetic type using case endings exclusively (as in Proto-Slavic and the archaic Slavic languages) is evident in the gradual elimination of the use of the locative forms without prepositions. The end result of this development is seen in Bulgarian and Macedonian, in which noun declension has almost completely disappeared and has been replaced by syntactic combinations using prepositions (*na kniga* 'of a book, to a book'). In Serbo-Croatian and in the western part of the West Slavic area (Sorbian and Czech), the same tendency to lose some of the distinctions between cases is observed, but to a lesser degree. In the other West Slavic languages and in East Slavic, on the other hand, the old system of declension by case endings has been preserved

Preservation of the system of declension in spite of the large number of loanwords and other neologisms that have no case distinctions at all (*e.g.*, borrowed Russian nouns like *kino* 'cinema,' or acronyms ending in a vowel like *Rayono* 'district education department').

Noun forms. The declension of pronouns has been preserved in all Slavic languages. Old combinations of adjectives with pronouns gave rise to the definite forms of adjectives (*e.g.*, feminine *dobra-ja* 'good-the'). These forms still contrast with the indefinite forms in South Slavic, but in the other languages the indefinite forms either have been gradually lost or else have been preserved only to serve a special function, that of predicate after 'to be.' In Bulgarian and Macedonian, as well as in some northern East Slavic dialects, an article is used, placed after a noun or adjective (*e.g.*, in Bulgarian and Macedonian, *kniga-ta* 'book-the,' *dobra-ta kniga* 'good-the book'). The three main genders are masculine, feminine, and neuter. Most Slavic languages distinguish animate and inanimate masculine noun forms; some (*e.g.*, Polish) also have personal and nonpersonal masculine forms.

Verb tenses. In the modern Slavic languages the verb is inflected to show present and past tenses. In the early history of the individual languages, however, a distinction was made between two past tenses, the aorist and the imperfect (the aorist denotes the occurrence of an action without reference to its completion, repetition, or duration; the imperfect is a verb tense designating a continuing state or an uncompleted action, especially in the past); this distinction is still preserved in modern South Slavic (with the exception of Slovene). Slavic has almost no traces of the Indo-European old perfect tense but, from combinations of a participle (verb + suffix *l* + masculine, feminine, or neuter endings) and forms of 'to be,' created new perfect (and pluperfect) tenses. Thus, from **dati* 'to give' there is a form **dalŭ jesmĭ* 'I have given' for a male speaker, **dala jesmĭ* for a female. Later these perfect forms came to be used as past tense forms in different Slavic languages. Slavic verbs usually come in pairs, one of which expresses the perfective (completed) and the other the imperfective (uncompleted) aspects of the same verb—*e.g.*, Russian *dat'* 'to give' (*i.e.*, 'to complete the process of giving'), *davat'* 'to be in the process of giving.'

The present tense form of a perfective verb may be used to express future meaning in East and West Slavic. Imperfective verbs need an auxiliary to make their future tense. South Slavic future tenses use an auxiliary (mostly from 'want') in both aspects. The eastern South Slavic languages, Bulgarian and Macedonian, have lost the infinitive form of the verb through the influence of non-Slavic Balkan languages, and they have developed verb forms to differentiate between an action witnessed by the speaker and one not witnessed (hence only reported).

Syntax. A striking feature of Slavic syntax is the widespread use of possessive adjectives (*e.g.*, Russian *Bož'ja milost'* 'divine mercy') instead of the genitive case of the noun (*milost' Boga* 'the mercy of God'). Word order in the Slavic languages is characterized by a gradual shift of the verb from the beginning to the middle of the sentence (subject–verb–object). Other important features of Slavic syntax are related to this medial positioning of the verb and the consequent occurrence of the verb before the object. For example, modifiers and prepositions are usually placed before nouns; today they follow nouns only in some set phrases like Church Slavonic *Boga radi* 'for God's sake,' with *radi* 'for the sake of' following the noun *Boga* 'God.'

Originally the verb occupied the initial position, which throws light on the origin of the reflexive verbal forms; these may be traced to the Proto-Slavic combination of the verb with a reflexive pronoun that occurred immediately after the verb and was pronounced as one accentual unit with the verb.

The rules for the shift of the stress in syntactic combinations with enclitics (an enclitic is a word treated in pronunciation as part of the preceding word) were identical for verbs and nouns. Depending on the accentuation of the verb or noun, the stress could be shifted either to the enclitic (as in Bulgarian *esen-és* 'last autumn') or to the proclitic, or preceding unstressed word (as *u* in Serbo-Croatian *ȕ jesēn* 'in the autumn').

Vocabulary. The original vocabulary of general terms common to Baltic and Slavic is still retained in most of the Slavic languages. In prehistoric times Proto-Slavic borrowed a number of important social and religious terms from Iranian (*e.g.*, *bogŭ* 'god,' *mirŭ* 'peace'). Later, spe-

Table 32: The Russian Alphabet

Cyrillic letters				equivalent		approximate pronunciation
printed		written		EB preferred	International Organization for Standardization	
capital	lower-case	capital	lower-case			
А	а	А	а	a		*f*a*ther*
Б	б	Б	б	b		*baby*
В	в	В	в	v		*vivid*
Г	г	Г	г	g		*go**
Д	д	Д	д, g	d		*did*
Е	е	Е	е	e or ye†	e	*bet* or *yet*
Ё	ё	Ё	ё	o or yo‡	ë	*yore*
Ж	ж	Ж	ж	zh	ž	*azure*
З	з	З	з, z	z		*zone*
И	и	И	и	i‖		*machine*
Й	й	Й	й	y‖	j	*boy*
К	к	К	к	k		*kin*
Л	л	Л	л	l		*lily*
М	м	М	м	m		*maim*
Н	н	Н	н	n		*no*
О	о	О	о	o		*order*
П	п	П	п	p		*pepper*
Р	р	Р	р	r		*error* (trilled)
С	с	С	с	s		*sand*
Т	т	Т	т	t		*tie*
У	у	У	у	u		*rule*
Ф	ф	Ф	ф	f		*fifty*
Х	х	Х	х	kh	x	Ger. Bu*ch*
Ц	ц	Ц	ц	ts	c	*cats*
Ч	ч	Ч	ч	ch	č	*chin*
Ш	ш	Ш	ш	sh	š	*shy*
Щ	щ	Щ	щ	shch	šč	ra*sh ch*oice or fre*sh sh*eet
Ъ	ъ		ъ	omit	††	
Ы	ы		ы	y‖		rh*y*thm
Ь	ь		ь	omit	**	
Э	э	Э	э	e	è (~ė)	*echo*
Ю	ю	Ю	ю	yu	ju	*you*th
Я	я	Я	я	ya	ja	*ya*rd

*Pronounced as *v* in genitive endings *-ego* and *-ogo*. †*e* after consonant; *ye* initially or after a vowel, ъ, or ь. ‡*o* after ж, ч, ш, щ; *yo* elsewhere. ‖The adjective endings -ИЙ and -ЫЙ can be transliterated simply *-y*; in linguistic discussion they are rendered as *-ij* and *-yj*. ††In linguistic discussion, rendered as close quotation marks ("). This symbol is the "hard sign"; it hardens the preceding consonant or separates syllables. **In linguistic discussion, rendered as an apostrophe ('). This symbol is the "soft sign"; it softens or palatalizes the preceding consonant or separates syllables.

Table 33: The Serbo-Croatian Alphabets

letters				letters			
Croatian		Serbian††		Croatian		Serbian††	
capital	lower-case	capital	lower-case	capital	lower-case	capital	lower-case
A	a	А	а	L	l	Л	л
B	b	Б	б	Lj	lj	Љ	љ
C	c	Ц	ц	M	m	М	м
Č	č	Ч	ч	N	n	Н	н
Ć	ć	Ћ	ћ	Nj	nj	Њ	њ
D	d	Д	д	O	o	О	о
Dž*	dž*	Џ	џ	P	p	П	п
Đ†	đ†	Ђ	ђ	R	r	Р	р
E	e	Е	е	S	s	С	с
F	f	Ф	ф	Š	š	Ш	ш
G	g	Г	г	T	t	Т	т
H	h	Х	х	U	u	У	у
I	i	И	и	V	v	В	в
J	j	Ј	ј	Z	z	З	з
K	k	К	к	Ž	ž	Ж	ж

*Alphabetized in *Britannica* as *dz*. †Alternatively, *dj*. ††The order of the Serbian Cyrillic letters is as follows:
А Б В Г Д Ђ Е Ж З И Ј К Л М Н Њ О П Р С Т Ћ У Ф Х Ц Ч Џ Ш.

cial terms were borrowed by East Slavic and South Slavic from eastern languages (especially Turkish) as a result of the political domination of the Tatars in Russia and of the Turks in the Balkans. After the Renaissance, loanwords were taken from classical and western European languages (especially German and French) into all the Slavic languages. Church Slavonic in its different variants remained the main source of innovations in vocabulary in East Slavic and in some South Slavic languages.

Extensive use of affixes

The Slavic languages make extensive use of prefixes and suffixes to derive new words and thereby enrich the vocabulary—*e.g.*, Russian *čern-yj* 'black,' *čern-i-t'* 'to blacken,' *o-čern-i-t'* 'to slander.' Several prefixes may be combined to modify the meaning of a verb (*e.g.*, Bulgarian *iz-po-raz-boleja se,* in which the added prefixes intensify the meaning 'for many people to fall ill'). Many derivational suffixes are common to most Slavic languages—*e.g.*, the very productive suffix *-stvo* (as in Russian *khristian-stvo* 'Christianity,' Ukrainian *pobratym-stvo* 'fraternity,' Polish *głup-stwo* 'foolishness, trifle,' Macedonian *golem-stvo* 'high status, arrogance').

The archaic type of derivation by compounding, inherited from Indo-European, was particularly productive in Church Slavonic under the stimulus of Greek. Compounding remains one of the methods of creating new terms, especially technical terms (*e.g.*, Russian *vodokhranilishche* 'reservoir' from *voda* 'water' and *khranilishche* 'depository'), but is far less important than affixation. Some Slavic languages typically derive new words by means of a condensed suffixing (*e.g.*, Czech *železnice* 'railroad,' from *železo* 'iron' combined with a noun-forming suffix; *hledisko* 'point of view,' from *hled* 'look' combined with a noun-forming suffix), while others tend to use combinations of words (*e.g.*, Russian *železnaja doroga* 'iron road' combined with an adjective-forming suffix on the first word; *točka zrenija* 'point of viewing').

Table 34: The Bulgarian Alphabet

Cyrillic letters: printed, capital	printed, lower-case	written, capital	written, lower-case	equivalent	approximate pronunciation
А	а	*А*	*а*	a	father
Б	б	*Б*	*б*	b	*baby*
В	в	*В*	*в*	v	*vivid*
Г	г	*Г*	*г*	g	*go*
Д	д	*Д*	*д, g*	d	*did*
Е	е	*Е*	*е*	e	bet
Ж	ж	*Ж*	*ж*	zh	azure
З	з	*З*	*з*	z	zone
И	и	*И*	*и*	i	machine
Й	й	*Й*	*й*	y	boy
К	к	*К*	*к*	k	*kin*
Л	л	*Л*	*л*	l	*lily*
М	м	*М*	*м*	m	*maim*
Н	н	*Н*	*н*	n	*no*
О	о	*О*	*о*	o	order
П	п	*П*	*п*	p	*pepper*
Р	р	*Р*	*р*	r	error (trilled)
С	с	*С*	*с*	s	sand
Т	т	*Т*	*т*	t	tie
У	у	*У*	*у*	u	rule
Ф	ф	*Ф*	*ф*	f	*fifty*
Х	х	*Х*	*х*	kh	Ger. Buch
Ц	ц	*Ц*	*ц*	ts	cats
Ч	ч	*Ч*	*ч*	ch	*chin*
Ш	ш	*Ш*	*ш*	sh	*shy*
Щ	щ	*Щ*	*щ*	sht	*shtick*
Ъ	ъ		*ъ*	ŭ*	above
Ь	ь		*ь*		†‡
Ю	ю	*Ю*	*ю*	yu	youth‡
Я	я	*Я*	*я*	ya	yard‡

*ŭ can occur initially in only a few words. †When transliterated, this is represented by an apostrophe. In modern orthography ’ is used only before *o*. ‡When *ya, yu,* and *’o* are preceded by consonants they are themselves pronounced like *a, u,* and *o*, respectively, but cause the preceding consonant to be palatalized, with a *y*-like pronunciation.

Writing systems. The first writing system used for Slavic was the Glagolitic system invented by St. Cyril. Quite original in pattern, it reflected accurately the sound system of the Macedonian dialect. Some forms of its letters can be traced to several different alphabets, mainly Greek and Semitic ones. Glagolitic was widely used in the first three centuries of Slavic literature but was gradually replaced by the Cyrillic alphabet, created in the 10th century and still used to write all the East Slavic languages, Bulgarian, Macedonian, and Serbian. Several languages (Serbian in the 19th century, Russian and Bulgarian in the 20th) have undergone reforms, dropping superfluous letters from the Cyrillic alphabet.

Other Slavic languages use the Latin (roman) alphabet. To render the distinctive sounds of a Slavic language, Latin letters are combined or diacritic signs are used (*e.g.*, Polish *sz* for the *sh* sound in *ship,* Czech *č* for the *ch* sound in *church*). An orthographic system devised by the Czech religious reformer Jan Hus (*c.* 1370–1415) was adopted into different West Slavic systems of writing, including Czech, Slovak, and Sorbian. Polish spelling was patterned after the pre-Hus Czech spelling of the 14th century. Most of the Slavic writing systems are constructed to symbolize the distinctive sounds of the language or to render the same morphemes by the same groups of letters despite differences in pronunciation in various forms. Modern Russian spelling reflects a morpheme-based principle.

(V.V.I./W.Br.)

Albanian language

Albanian is an Indo-European language spoken by some five million inhabitants of the eastern Adriatic coast in Albania and also in neighbouring Yugoslavia (principally in Kosovo) and Macedonia, west of a line from near Leskovac (Yugoslavia) to Lake Ohrid. There are perhaps 200,000 more speakers in isolated villages in southern Italy (in Abruzzi, Molise, Basilicata, Puglia, and Calabria), and Sicily, and southern Greece (in Boeotia [Voiotía], Attikí, Euboea [Évvoia], and the Peloponnese [Pelopónnisos]), and on the island of Ándros.

The origins of the general name Albanian, which traditionally referred to a restricted area in central Albania, and of the current official name Shqip or Shqipëri, which may well be derived from a term meaning "pronounce clearly, intelligibly," are still disputed. The name Albanian has been found in records since the time of Ptolemy. In Calabrian Albanian the name is Arbresh, in Modern Greek Arvanítis, and in Turkish Arnaut; the name must have been transmitted early through Greek speech.

Gheg and Tosk dialects

Dialects. The two principal dialects, Gheg in the north and Tosk in the south, are separated roughly by the Shkumbin River. Gheg and Tosk have been diverging for at least a millennium, and their less extreme forms are mutually intelligible. Gheg has the more marked subvarieties, the most striking of which are the northernmost and eastern types, which include those of the city of Shkodër (Scutari), the northeastern Skopska Crna Gora region of Macedonia, Kosovo (in Yugoslavia), and the isolated village of Arbanasi (outside Zadar) on the Croatian coast of Dalmatia. Arbanasi, founded in the early 18th century by refugees from the region around the Montenegrin coastal city of Bar, has about 2,000 speakers.

All of the Albanian dialects spoken in Italian and Greek enclaves are of the Tosk variety and seem to be related most closely to the dialect of Çamëria in the extreme south of Albania. These dialects resulted from incompletely understood population movements of the 13th and 15th centuries. The Italian enclaves—nearly 50 scattered villages—probably were founded by emigrants from Turkish rule in Greece. A few isolated outlying dialects of south Tosk origin are spoken in Bulgaria and Turkish Thrace but are of unclear date. The language is still in use in Mandritsa, Bulgaria, at the border near Edirne, and in an offshoot of this village surviving in Mándres, near Kilkís in Greece, that dates from the Balkan Wars. A Tosk enclave near Melitopol in Ukraine appears to be of moderately recent settlement from Bulgaria. The Albanian dialects of Istria, for which a text exists, and of Syrmia (Srem), for which there is none, have become extinct.

History. The official language, written in a standard roman-style orthography adopted in 1909, was based on the south Gheg dialect of Elbasan from the beginning of the Albanian state until World War II and since has been modelled on Tosk. In Yugoslavia, Albanian speakers in the region of Kosovo in Serbia (officially bilingual in Serbian and Albanian) and in Macedonia speak eastern varieties of Gheg but since 1974 have widely adopted a common orthography with Albania. Before 1909 the little literature that was preserved was written in local makeshift Italianate or Hellenizing orthographies or even in Turko-Arabic characters.

The first written records

A few brief written records are preserved from the 15th century, the first being a baptismal formula from 1462. The scattering of books produced in the 16th and 17th centuries originated largely in the Gheg area (often in Scutarene north Gheg) and reflect Roman Catholic missionary activities. Much of the small stream of literature in the 19th century was produced by exiles. Perhaps the earliest purely literary work of any extent is the 18th-century poetry of Gjul Variboba, of the enclave at S. Giorgio, in Calabria. Some literary production continued through the 19th century in the Italian enclaves, but no similar activity is recorded in the Greek areas. All these early historical documents show a language that differs little from the current language. Because these documents from different regions and times exhibit marked dialect peculiarities, however, they often have a value for linguistic study that greatly outweighs their literary merit.

Classification. That Albanian is of clearly Indo-European origin was recognized by the German philologist Franz Bopp in 1854; the details of the main correspondences of Albanian with Indo-European languages were elaborated by another German philologist, Gustav Meyer, in the 1880s and '90s. Further linguistic refinements were presented by the Danish linguist Holger Pedersen and the Austrian Norbert Jokl. The following etymologies illustrate the relationship of Albanian to Indo-European (an asterisk preceding a word denotes an unattested, hypothetical Indo-European parent word, which is written in a conventionalized orthography): *pesë* "five" (from **pénk^{w}e*); *zjarm* "fire" (from **g^{w}hermos*); *natë* "night" (from **nok^{w}t-*); *dhëndër* "son-in-law" (from **ǵemə ter-*); *gjarpër* "snake" (from **sérpŏ̄n-*); *bjer* "bring!" (from **bhere*); *djeg* "I burn" (from **dheg^{w}hō*); *kam* "I have" (from **kapmi*); *pata* "I had" (from **pot-*); *pjek* "I roast" (from **pek^{w}ō*); *thom, thotë* "I say, he says" (from **k'ēmi, *k'ēt* . . .).

The verb system includes many archaic traits, such as the retention of distinct active and middle personal endings (as in Greek) and the change of a stem vowel *e* in the present to *o* (from **ē*) in the past tense, a feature shared with the Baltic languages. For example, there is *mbledh* "gathers (transitive)" as well as *mblidhet* "gathers (intransitive), is gathered" in the present tense, and *mblodha* "I gathered" with an *o* in the past. Because of the superficial changes in the phonetic shape of the language over 2,000 years and because of the borrowing of words from neighbouring cultures, the continuity of the Indo-European heritage in Albanian has been underrated.

Albanian shows no obvious close affinity to any other Indo-European language; it is plainly the sole modern survivor of its own subgroup. It seems likely, however, that in very early times the Balto-Slavic group was its nearest of kin. Of ancient languages, both Dacian (or Daco-Mysian) and Illyrian have been tentatively considered its ancestor or nearest relative.

Grammar. The grammatical categories of Albanian are much like those of other European languages. Nouns show overt gender, number, and three or four cases. An unusual feature is that nouns are further inflected obligatorily with suffixes to show definite or indefinite meaning: *e.g., bukë* "bread," *buka* "the bread." Adjectives—except numerals and certain quantifying expressions—and dependent nouns follow the noun they modify; and they are remarkable in requiring a particle preceding them that agrees with the noun. Thus, in *një burrë i madh,* meaning "a big man," *burrë* "man" is modified by *madh* "big," which is preceded by *i,* which agrees with the term for "man"; likewise, in *dy burra të mëdhenj* "two big men," *mëdhenj,* the plural masculine form for "big," follows the noun *burra* "men" and is preceded by a particle *të* that agrees with the noun. Verbs have roughly the number and variety of forms found in French or Italian and are quite irregular in forming their stems. Noun plurals are also notable for the irregularity of a large number of them. When a definite noun or one taken as already known is the direct object of the sentence, a pronoun in the objective case that repeats this information must also be inserted in the verb phrase; *e.g., i-a dhashë librin atij* is literally "him-it I-gave the-book to-him," which in standard English would be "I gave the book to him." In general, the grammar and formal distinctions of Albanian are reminiscent of Modern Greek and the Romance languages, especially of Romanian. The sounds suggest Hungarian or Greek, but Gheg with its nasal vowels strikes the ear as distinctive.

Distinctive sound

Vocabulary and contacts. Although Albanian has a host of borrowings from its neighbours, it shows exceedingly few evidences of contact with ancient Greek; one such is the Gheg *mokën* (Tosk *mokër*) "millstone," from the Greek *mēkhanḗ*. Obviously close contacts with the Romans gave many Latin loans—*e.g., mik* "friend" from Latin *amicus; këndoj* "sing, read" from *cantāre.* Furthermore, such loanwords in Albanian attest to the similarities in development of the Latin spoken in the Balkans and of Romanian, a Balkan Romance tongue. For example, Latin *palūdem* "swamp" became *padūlem* and then *pădure* in Romanian and *pyll* in Albanian, both with a modified meaning, "forest."

Conversely, Romanian also shares some apparently non-Latin indigenous terms with Albanian—*e.g.,* Romanian *brad,* Albanian *bredh* "fir." Thus these two languages reflect special historical contacts of early date. Early communication with the Goths presumably contributed *tirq* "trousers, breeches" (from an old compound "thigh-breech"), while early Slavic contacts gave *gozhdë* "nail." Many Italian, Turkish, Modern Greek, Serbian, and Macedonian-Slav loans can be attributed to cultural contacts of the past 500 years with Venetians, Ottomans, Greeks (to the south), and Slavs (to the east).

A fair number of features—*e.g.,* the formation of the future tense and of the noun phrase—are shared with other languages of the Balkans but are of obscure origin and development; Albanian or its earlier kin could easily be the source for at least some of these. The study of such regional features in the Balkans has become a classic case for research on the phenomena of linguistic diffusion.

(E.P.H.)

URALIC LANGUAGES

Uralic languages are spoken by more than 21,000,000 people scattered throughout northeastern Europe, northern Asia, and (through immigration) North America. The language family consists of some 21 languages, the most demographically important of which is Hungarian, the official language of Hungary.

Attempts to trace the genealogy of the Uralic languages to periods earlier than Proto-Uralic have been hampered by the great changes in the attested languages, which preserve relatively few features and therefore provide little evidence upon which scholars may base meaningful claims for a more distant relationship. Most commonly mentioned in this respect is a putative connection with the Altaic language family (including Turkic and Mongolian). This hypothetical language group, called Ural-Altaic, is not considered by most scholars to be soundly based. Although the Uralic and Indo-European languages are not generally thought to be related, more specula-

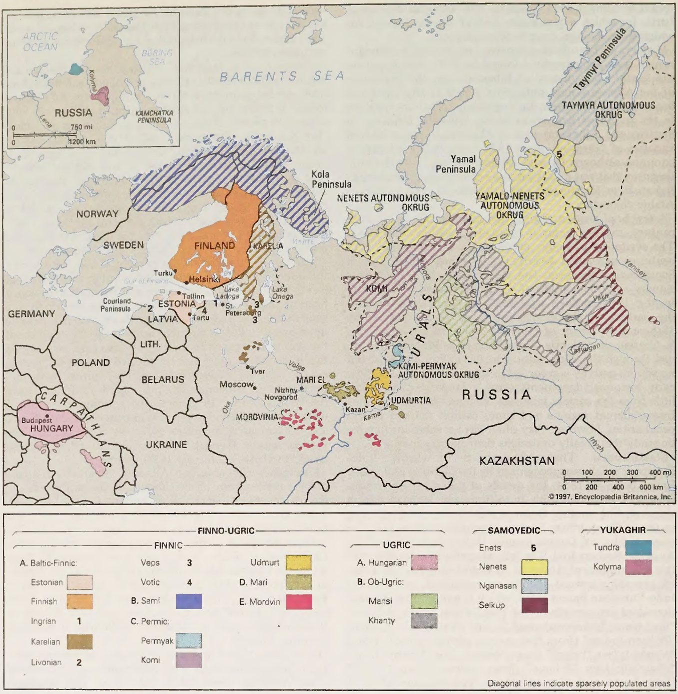


Figure 17: Distribution of the Uralic languages.

tive studies have suggested a connection between them. Relationship with the Eskimo languages, Dravidian (*e.g.,* Telugu), Japanese, Korean, and various American Indian groups has also been proposed. The most radical of these claims is the massive Dené-Finnish grouping of Morris Swadesh, which encompasses, among others, Sino-Tibetan (*e.g.,* Chinese) and Athabaskan (*e.g.,* Navajo).

The Uralic language family in its current status consists of two related groups of languages, the Finno-Ugric and the Samoyedic, both of which developed from a common ancestor, called Proto-Uralic, that was spoken 7,000 to 10,000 years ago in the general area of the north-central Ural Mountains. At its very earliest stages Uralic most probably included the ancestors of the Yukaghir languages (formerly listed as a Paleo-Siberian stock with no known relatives).

Over the millennia, both Finno-Ugric and Samoyedic branches of Uralic have given rise to more or less divergent subgroups of languages, which nonetheless have retained certain traits from their common source. For example, the degree of similarity between two of the least closely related members of the Finno-Ugric group, Hungarian and Finnish, is comparable to that between English and Russian (which belong to the Indo-European family of languages). The difference between any Finno-Ugric language and any Samoyedic tongue would be even greater. On the other hand, more closely related members of Finno-Ugric, such as Finnish and Estonian, differ in much the same manner as greatly diverse dialects of the same language.

The distribution of Uralic languages

ESTABLISHMENT OF THE FAMILY

Determining the geographic location, material culture, and linguistic characteristics of the earliest stages of Uralic at a period thousands of years prior to any historical record is a problem beset with enormous difficulties; consensus among Uralic scholars is limited to a handful of general hypotheses.

The homeland of Proto-Uralic

The original homeland of Proto-Uralic is considered to have been in the vicinity of the north-central Urals, possibly centred west of the mountains. Following the dissolution of Uralic, the precursors of the Samoyeds gradually moved northward and eastward into Siberia. The Finno-Ugrians moved to the south and west, to an area close to the confluence of the Kama and Volga rivers.

Shared cognates. Several kinds of indirect evidence support the above supposition. One approach attempts to reconstruct the natural environment of these groups on the basis of shared cognates (related words) for plants, animals, and minerals and on the distribution of these words in the modern languages. For example, cognates designating certain types of spruce are found in all the Uralic languages except Hungarian (Finnish *kuusi,* Sami [Lapp] *guossâ,* Mordvin *kuz,* Komi *koz,* Khanty *kol,* Nenets *xādy,* Selkup *kūt*). Because the range of this type of fir tree is restricted to more northern climates, it is generally assumed that the widespread consistent association of the name and the tree suggests a period in which Proto-Uralic was spoken within that zone. Several other terms for plants (*e.g.,* Finnish *muurain* "cloudberry" [*Rubus arcticus*]), a term for metal (Estonian *vask* "copper," Hungarian *vas* "iron," Nganasan *basa* "iron"), and a word for "reindeer" (Sami *boaʒo*) are also consistent with a northern Ural location. Great caution is necessary in such matters, because the association of words and objects also can result from borrowing, perhaps long after the period of Uralic unity; especially such culturally mobile items as "metal" and "reindeer" cannot be traced with certainty to a Proto-Uralic community. The central Volga location of Proto-Finno-Ugric is strongly supported by an abundance of shared terminology dealing with beekeeping, which constitutes a significant part of the culture of this region.

Contacts with unrelated languages. A second approach to determining the location of Proto-Uralic is based on contacts with other, unrelated languages as evidenced by loanwords from one group to the other. Early Finno-Ugric borrowed numerous terms from very early dialects of Indo-European. Though these words are entirely lacking from the Samoyed languages, within the Finno-Ugric division they are shared by the most remotely related members and show the same phonetic relationships as the native Finno-Ugric vocabulary. Examples include agricultural and apicultural terminology (*e.g.,* "honey": Finnish *mete,* Komi *ma,* Hungarian *méz* [compare Indo-European **medhu*-]; "pig": Finnish *porsas,* Komi *porś*); several numerals ("hundred": Finnish *sata,* Hungarian *száz*); mineral words ("salt": Finnish *suola,* Komi *sol*); and the word for orphan (Finnish *orpo,* Hungarian *árva*). The nature of these borrowings, together with the linguist's relatively richer knowledge of early Indo-European, supports a southward movement of Proto-Finno-Ugric and also provides some insight into the culture of the Finno-Ugrians.

The distribution of the daughter languages. The central Volga origin hypothesis is also supported by the geographic distribution of the daughter languages. Except for Hungarian, which moved westward across the steppes, the Finno-Ugric languages form two chains distributed along major waterways, with the confluence of the Kama and Volga at their centre. One chain extends northward along the Kama, across the northern tip of the Urals into the Ob watershed, then southward along the Ob and its tributaries. The second extends to the northeast along the Volga to the Gulf of Finland. The extinct Merya, Murom, and Meshcher languages were once links in this chain. Finally, assumptions about the more distant relationships of Uralic have influenced views concerning its original location. Earlier, proponents of the Ural-Altaic hypothesis tended to place the Uralic homeland in south-central Siberia, near the sources of the Ob and the Yenisey, but there is no substantive support for this view.

Two chains of daughter languages

CURRENT DISTRIBUTION

The Finno-Ugric languages are represented today by some 20 languages scattered over an immense Eurasian territory. In the west they include the European national languages Hungarian, Finnish, and Estonian as well as the Sami (or Lapp) languages, the westernmost members of the group, spoken by numerous distinct communities across the northern Scandinavian Peninsula from central Norway to the White Sea. The remaining Finno-Ugric languages are located in the Baltic countries and in Russia, all formerly republics of the Soviet Union, with one major concentration—which includes Estonian, Livonian, Votic, Karelian, and Veps—extending from the Gulf of Riga to the Kola Peninsula. The Mordvin and Mari languages are found in the central Volga region; from there extending northward along river courses west of the Urals are the Permic languages—Udmurt, Komi (Zyryan), and Permyak (or Komi-Permyak). East of the Urals, along the Ob River and its tributaries, are the easternmost representatives of the Finno-Ugric group—Mansi and Khanty.

The largely nomadic Samoyeds are sparsely distributed over an enormous area extending inward from the Arctic shores of Russia from the White Sea in the west to Khatanga Bay in central Siberia in the east. Nenets, the westernmost of these languages, reaches eastward to the mouth of the Yenisey River and includes a small insular group on Novaya Zemlya. Speakers of Enets are located in the region of the upper Yenisey. The lower half of the Taymyr Peninsula is the habitat of the Nganasan, the easternmost of the Uralic groups. The fourth language, Selkup, lies to the south in a region between the central Ob and central Yenisey; its major representation is located between Turukhansk and the Taz River. A fifth Samoyedic language, Kamas (Sayan), spoken in the vicinity of the Sayan Mountains, survived into the 20th century but is now extinct. Yukaghir is represented by two small language groups (designated Tundra and Kolyma) in far northeastern Siberia, between the tundra east of the Alazeya River and the upper tributaries of the Kolyma.

The political history of the various Uralic groups largely has been one of resisting encroachment from adjacent European (especially Germanic and Slavic) and Turkic groups and from other Uralic neighbours. Only the three largest and westernmost groups have succeeded in achieving political independence—Hungary, Finland, and Estonia. The political status of the Uralic groups within Russia generally reflects their demographic significance. The five largest minority groups, with populations ranging from 100,000 to almost 1,000,000 speakers, are centred in the largely autonomous republics of Mordvinia, Mari El, Udmurtia, Komi, and Karelia. Four other groups possess autonomy to a lesser degree: the Khanty and the Mansi (in Khanty-Mansi autonomous *okrug*), the speakers of Permyak (in Komi-Permyak autonomous *okrug*), and the Nenets (in Taymyr, Nenets, and Yamalo-Nenets *okrugs*). The Sami, who are widely distributed across four countries (Norway, Sweden, Finland, and Russia), have achieved only local political recognition. A number of the smaller Uralic language communities, such as Livonian and Votic, face extinction through cultural assimilation by the end of the century.

Minority Uralic languages in Russia

Because the names designating many of the Uralic peoples have never been standardized, a wide range of appellations is encountered in references to these groups. Earlier designations, especially in the case of the groups in Russia, tended to be taken from derogatory names used by neighbouring peoples—*e.g.,* Cheremis, now Mari. Table 35 indicates the names in use. Standard usage is in the left column, and earlier, Russian-based forms are in parentheses. The name that the group uses for itself and certain other information, such as Russian and Old Russian forms, are in the right column. Several names are identical to the word for "man" in these languages. (Finnish *mies* "man" also has been etymologically related to the names Magyar and Mansi.) It is important that Khanty (Ostyak) be differentiated from Selkup (Ostyak Samoyed) and from Ket (Yenisey Ostyak, a non-Uralic tongue), which should not be confused with Enets (Yenisey).

Table 35: Names Used to Designate Uralic Groups

English form	native form
Finnish	*suomi*
Karelian	*karjala*
Ingrian	*izhor*
Veps	*vepsä, lüüd* (Old Russian *vesj, chudj*)
Estonian	*eesti* (Old Russian *chudj*)
Votic, Vote	*vadja* (Old Russian *vodj; chudj*)
Livonian	*liiv*
Sami (Lapp)	*sabme* (Russian *saami;* earlier *lopj*)
Mordvin	*erza, moksha* (no common name)
Mari (Cheremis)	*mari* ("man")
Udmurt (Votyak)	*ud-murt* (*murt* = "man")
Komi (Zyryan)	*komi* (Old Russian *permj*)
Khanty (Ostyak)	*khanty* (Old Russian *jugra*)
Mansi (Vogul)	*manshi* (also designates the Khanty; Old Russian *jugra*)
Hungarian	*magyar* (Russian *vengr*)
Nenets (Yurak)	*nenets, hasawa* ("man"; Old Russian *samojadj*)
Enets (Yenisey)	*enetj* (related to the name *nenets*)
Nganasan (Tavgi, Avam)	*ŋanasan* (related to the name *nenets*)
Selkup (Ostyak Samoyed)	*shöl-qup* (*shö*[*l*] = "earth," *qup* = "man")
Yukaghir	*wadu, odu*

Languages of the family

The two major branches of Uralic are themselves composed of numerous subgroupings of member languages on the basis of closeness of linguistic relationship. Finno-Ugric can first be divided into the most distantly related Ugric and Finnic (sometimes called Volga-Finnic) groups, which may have separated as long ago as five millennia. Within these, three relatively closely related groups of languages are found: the Baltic-Finnic, the Permic, and the Ob-Ugric. The largest of these, the Baltic-Finnic group, is composed of Finnish, Estonian, Livonian, Votic, Ingrian, Karelian, and Veps. The Permic group consists of Komi, Permyak, and Udmurt; the Ob-Ugric group includes Mansi and Khanty.

Divisions within Finno-Ugric

The Ugric group comprises the geographically most distant members of the family—the Hungarian and Ob-Ugric languages. Finnic contains the remaining languages: the Baltic-Finnic languages, the Sami (or Lapp) languages, Mordvin, Mari, and the Permic tongues. There is little accord on the further subclassification of the Finnic languages, although the fairly close relationship between Baltic-Finnic and Sami is generally recognized (and is called North Finnic); the degree of separation between the two may be compared to that between English and German. Mordvin has most frequently been linked with Mari (a putative Volga language group), but comparative evidence also suggests a bond with Baltic-Finnic and Sami (that is, West Finnic). The extinct Merya, Murom, and Meshcher tongues, known only from Old Russian chronicles, are assumed to have been spoken by Finnic peoples and, from their geographic location northwest of Mordvin, must have belonged to West Finnic. One hypothesis for the internal relationships of the Uralic family as a whole is given in Figure 18.

The precursor of the modern Samoyedic languages is thought to have divided near the beginning of the 1st century AD into a northern and a southern group. North Samoyedic consists of Nenets, Enets, and Nganasan. South Samoyedic contains a single living language, Selkup, and numerous other dialects now extinct: Kamas, Motor, Koibal, Karagas, Soyot, and Taigi.

UGRIC

Hungarian. Hungarian, the official language of Hungary, remains the primary language of the fertile Carpathian Basin. Bounded by the Carpathian Mountains to the north, east, and southwest, the Hungarian language area is represented by several million speakers outside the boundaries of Hungary—mostly in Romanian Transylvania and in Slovakia. To the south a substantial Hungarian population extends into Croatia and Yugoslavia. Hungarian emigrant communities are found in many parts of the world, especially in North America and Australia.

The ancestors of the Hungarians, following their separation from the other Ugric tribes, moved south into the steppe region below the Urals. As mounted nomads, in contact with and often in alliance with Turkic tribes, they moved westward, reaching and conquering the sparsely settled Carpathian Basin in the period 895–896. The Hungarians came under the influence of Rome through their first Christian king, Stephen (István), in 1001, and the use of Latin for official purposes continued into the 19th century. Following a Hungarian defeat at the Battle of Mohács in 1526, Hungary was occupied by Turkish forces, who were replaced by German Habsburg domination in the late 17th century. Concern for a common literary medium, closely tied with Hungarian nationalism, began in the late 18th century. More recent foreign influences on the language were suppressed and replaced by native words and constructions. The literary form received a broad dialect base, facilitating its use as a national language.

Modern Hungarian has eight major dialects, which permit a high degree of mutual intelligibility. Budapest, the nation's capital, is located near the junction of three dialect areas: the South, Trans-Danubian, and Palóc (Northwestern). As a result of unfavourable treaties following both world wars, especially the Treaty of Trianon, two dialects (Central Transylvanian and Székely) lie almost entirely within Romania, and the remaining six dialects radiate outward into neighbouring countries.

Dialects of Hungarian

The Hungarians' own name for themselves is *magyar.* Other Western appellations, such as the French *hongrois,* German *Ungar,* and Russian *vengr,* all stem from the name of an early Turkic tribal confederation, the *on-ogur* (meaning "10 tribes"), which the Hungarians joined in their wanderings toward the west, and does not indicate relationship with the ancient Huns, a Turkic tribe. One of the earliest recorded references to the Hungarians, a Byzantine geographic survey of Constantine VII Porphyrogenitus (d. 959) entitled *De administrando imperio,* lists the *megyer* as one of the Hungarian tribes, but, as was typical in early reports, the Hungarians were not distinguished from their Turkish allies.

Ob-Ugric: Khanty and Mansi. Widely dispersed along the Ob River and its tributaries, the so-called Ob-Ugric

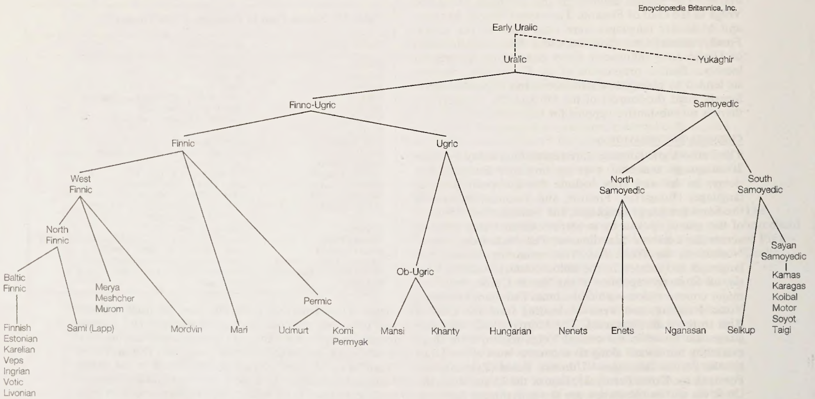


Figure 18: Family tree diagram of the Uralic languages, including their probable relationship to Yukaghir.

peoples, the Khanty and the Mansi, are among the least demographically significant of the Finno-Ugric groups. Although the Khanty have decreased in number over the past few centuries, their language is still maintained by about 14,000 speakers. The Mansi, by contrast, had only some 8,000 ethnic representatives by the end of the 20th century; of these, fewer than half were said to claim Mansi as their mother tongue. To a large extent both groups have been assimilated by their Russian and Tatar neighbours.

It is likely that the precursors of the Ob-Ugric tribes were still centred west of the Urals well within historic times, long after the division of Proto-Ugric into distinct languages. *The Russian Primary Chronicle* of Nestor, which assigned to the Khanty and Mansi the common name *jugra,* places them in the vicinity of the Pechora River in 1092; they did not shift to the Ob waterways until several centuries later.

Both groups live for the most part within the Khanty-Mansi autonomous *okrug* (district), which has its administrative centre in Khanty-Mansiysk at the confluence of the Ob and Irtysh rivers. The Khanty are concentrated along the Ob and its eastern tributaries, while the Mansi are found along the western tributaries primarily north of the Irtysh and just east of the Urals; a few Mansi speakers are also found in the Arctic lands west of the Urals.

Because of the great distances between the various groups, the dialects of both languages show considerable divergence. They are usually designated by the name of the river on which they are spoken. Mansi has four main dialect groups, of which one (Tavda) is practically extinct and another (Konda) is spoken only by individuals above a certain age. The largest dialect group (Northern) is centred on the Sosva and serves as the basis for the literary language. Khanty is divided into three main dialects: a northern dialect in the general area of the mouth of the Ob, an eastern dialect extending from east of the Irtysh to the Vakh and Vasyugan tributaries, and a southern dialect lying between the other two. Literary Khanty has been based primarily on the northern group, but standardization remains weak.

Ob-Ugric languages in print

Both of the Ob-Ugric languages first appeared in printed form in 1868 as a result of Gospel translations published in London, but it was not until after the formation of their autonomous *okrug* in 1930 that any sort of literary form of either language really existed. Until 1937 numerous books were published using a modified Latin (roman) alphabet; since then Cyrillic has been used. Some elementary education is conducted in the native languages within the *okrug.*

FINNIC

Finnish. Finnish, together with Swedish (an unrelated North Germanic language), serves as an official language of Finland. It is now spoken by more than 5,000,000 people, including about 95 percent of the inhabitants of Finland plus nearly 500,000 Finns in North America, Sweden, and Russia. It is also recognized as an official language in Russia's Karelian region, alongside Russian.

Finnish as the common language of the Finns is not the direct descendant of one of the original Baltic-Finnic dialects; rather, it arose through the interaction of several separate groups in the territory of modern Finland. These included the Häme; the southwestern Finns (originally called Suomi), who appear to be close relatives to the Estonians because they arrived directly from across the Gulf of Finland; and the Karelians, perhaps themselves a blend of Veps and more western Finnic groups. Early Russian chronicles refer to these as *jemj, sumj,* and *korela.* The intermixture of the three groups is still reflected in the distribution of the five main modern dialects, which form a western and an eastern area. The western area contains the southwestern dialect (near Turku), Häme (south-central), and a northern dialect subgroup (largely a mixture of the other two plus eastern traits). The eastern area consists of the Savo dialect (perhaps a blend of the original Karelian and Häme dialects) and a southeastern dialect, which strongly resembles Karelian. The Finnish word for their land and their language is *suomi,* the original meaning of which is uncertain. The first use of the term Finn (*fenni*) is found in the 1st century AD in Tacitus' *Germania,* but this usage is generally considered to refer to the ancestors of the Sami, who have also been labeled Finns at various times. (The province of Norwegian Lappland is called Finnmark.)

The first book in Finnish was an alphabet book from 1543 by Mikael Agricola, founder of the Finnish literary language; Agricola's translation of the New Testament appeared five years later. Finnish was accorded official status in 1809, when Finland entered the Russian Empire after six centuries of Swedish domination. The publication of the national folk epic, the *Kalevala,* created from folk songs collected among the eastern dialects by the folklorist and philologist Elias Lönnrot (first edition in 1835; substantially expanded in 1849), gave increased impetus to the movement to develop a common national language encompassing all dialect areas.

Estonian. Estonian serves as the official language of Estonia, located immediately south of Finland across the Gulf of Finland. Most of the some 1,000,000 speakers of Estonian live within Estonia, but others can be found in Russia, North America, and Sweden. Modern Estonian is the descendant of one or possibly two of the original Baltic-Finnic dialects. The modern language has two major dialects, a northern one, which is spoken in most of the country, and a southern one, which extends from Tartu to the south. The northernmost dialects share many features with the southwestern Finnish dialect. The Estonians' own name *eesti* came into general use only in the 19th century. The name *aestii* is first encountered in Tacitus, but it is likely that it referred to neighbouring Baltic-Finnic peoples.

Modern Estonian dialects

The first connected texts in Estonian are religious translations from 1524; the *Wanradt-Koell Catechism,* the first book, was printed in Wittenberg in 1535. Two centres of culture developed—Tallinn (formerly Revel) in the north and Tartu (Dorpat) in the south; in the 17th century each gave rise to a distinct literary language. Influenced by the Finnish *Kalevala,* the Estonian author F. Reinhold Kreutzwald fashioned a national epic, *Kalevipoeg* ("The Son of Kalevi"), which appeared in 20 songs between 1857 and 1861. As with the *Kalevala,* this was instrumental in kindling renewed interest in a common national literary language in the late 19th century.

Smaller Baltic-Finnic groups. The five less numerous Baltic-Finnic groups—Karelian, Veps, Ingrian, Votic, and Livonian—lie within Russia and the Baltic nations, largely in the general vicinity of the Gulf of Finland. The Karelians, Veps, and Livonians were among the original Baltic-Finnic tribes; Votic is considered to be an offshoot of Estonian, and Ingrian a remote branch of Karelian. None of these languages currently has a literary form, although unsuccessful initial attempts to establish one have been made for all but Votic (for Livonian as early as the 19th century, for the others during the 1930s). Since the beginning of the 20th century, the numbers of these Baltic-Finnic speakers have been drastically reduced, and, with the exception of Karelian and Veps, their extinction within several generations seems certain. Ingrian, Votic, and Livonian each have fewer than 1,000 speakers.

Karelian, the largest of these groups, with about 86,000 speakers—not counting those Karelians who emigrated into Finland following World War II—lies along a broad zone just east of the Finnish border from just north of St. Petersburg to the White Sea. A separate group of Karelians is found far to the south near Tver (formerly Kalinin) on the upper Volga. Karelian has two major dialects, Karelian proper and Olonets (*aunus* in Finnish), which is spoken northeast of Lake Ladoga. One of the first historical mentions of the Karelians is found in a report of the Viking Ohthere to King Alfred of England at the end of the 9th century; this indicates that they were already on the southern Kola Peninsula as neighbours of the Sami and gives their name as *beorma.*

The language of one of the original Baltic-Finnic tribes, Veps, is spoken southeast on a line connecting lower Lake Ladoga with central Lake Onega. Less than one-fifth of the ethnic population of some 14,000 Veps still consider the language their native tongue—a sharp decline from the 26,172 speakers reported in the mid-1800s. A small Baltic-Finnic group, composed of the Ludic dialects, is found between Veps and Karelian and is generally considered a blend of the two major groups rather than a separate language; the dialects are more closely akin to Karelian. The Ingrians and the Votes live on the southern Gulf of Finland in the border area between Estonia and Russia, where they survived because the border area was for many years closed to outsiders, even to visitors from other parts of the Soviet Union. Livonian has persisted in a dozen villages on the northernmost tip of Latvia, on the Courland Peninsula, but the language is not used by the younger generation.

SAMI AND OTHER FINNIC GROUPS

The Sami (Lapp) languages. The Sami are widely distributed, inhabiting territory from central Norway northward and eastward across northern Sweden and Finland to the Kola Peninsula. Their numbers have increased over the past century to more than 30,000, but the number of Sami speakers has declined rapidly since 1950 as the language has given way to the various official national languages. Sami is generally divided into three main dialect groups, each comprising various subtypes. These dialects are virtually mutually unintelligible, so that when speakers of different Sami groups meet they generally converse in Finnish, Swedish, or Norwegian. To speak of a single Sami (or Lapp) language is therefore misleading. Sami represents a group of at least four or five languages at least as diverse as the separate Baltic-Finnic languages. The largest group, North Sami (with approximately two-thirds of all speakers), is centred in northern Norway, Sweden, and Finland. East Sami consists of two small groups in eastern Finland—Inari and Skolt—in addition to Kola Sami in Russia. South Sami is still represented by a few speakers scattered from central Norway to north-central Sweden.

Mutual unintelligibility

North Sami has had a literary tradition that began with the 17th-century Swedish Sami Bible and other religious translations; in the mid-20th century elementary schools that used Sami as the language of instruction were found in many larger North Sami communities. Two basic variants of the literary language are in use. One, in Norway and Sweden, employs a special Sami orthographic system devised to accommodate a wide range of dialectal variation; a second, in Finland, is based on a narrower adaptation of Finnish orthography. Each of the two types has numerous local variants, and progress toward a common Sami orthography has been slow.

It is clear that the Sami were already present north of the Gulf of Finland prior to the arrival of the first Baltic-Finnic tribes, and from there they may have extended over much of the Scandinavian Peninsula.

They have been mentioned as the northern neighbours of the north Germanic tribes in numerous historical sources of the 1st millennium of the Christian Era. The Sami were taxed by the Norwegians in the 9th century and by the Karelians in the 13th century and since that time have continually retreated northward under pressure from their southern neighbours. The Sami's own name for themselves, *sabme,* is etymologically related to the Finnish dialect name, *häme*.

Other Finnic languages. Mordvin, Mari, and two of the Permic languages—Udmurt and Komi—are recognized by separate republics within Russia (respectively Mordvinia, Mari El, Udmurtia, and Komi). They also share official status with the Russian language. Mordvin, Mari, and Udmurt are centred on the middle Volga River, in roughly the area considered to have been the original home of Proto-Finno-Ugric. Because of their location, the history of these groups over the past millennium has been closely tied to that of the Turkic Bulgars, the Tatars (until 1552), and then the Russians. The Komi, having moved far to the north, eventually reaching into the Arctic tundra, did not come under Bulgar or Tatar influence. Old Permic, a written form of early Komi, was used in religious manuscripts in the 14th century, and a native Komi literary tradition stems from the 19th century. Grammars of Mari and Udmurt prepared by Russian linguists appeared in 1775, but native literary development in these languages, as well as in Mordvin, is of recent origin. Although these groups enjoyed the status of large minorities during the Soviet era, their numbers have increased over the past century, and they have maintained ethnic consciousness.

Continued ethnic consciousness

Mordvin. Mordvin, with more than 750,000 speakers (about two-thirds of the 1,153,000 Mordvins reported in 1989), is the fourth-largest Uralic group. The Mordvins are widely scattered over an area between the Oka and Volga rivers, some 200 miles southwest of Moscow. Less than half of their number live within the republic of Mordvinia. Mordvin has two main dialects, Moksha and Erzya, which are sometimes considered separate languages. Both have literary status. Although the Mordvins do not have a common designation for themselves beyond the two dialect names, the name Mordens appears in the 6th-century *Getica* of Jordanes and is no doubt related to the Permic word for "man," *murt/mort*.

Mari. Mari (formerly known as Cheremis) is currently maintained by about 610,000 speakers (approximately three-fourths of the ethnic Mari). They live primarily in an area north of the Volga between Kazan and Nizhny Novgorod, northeast of the Mordvin area, especially within Mari El republic. Mari El's three main dialects are the Meadow dialect, used by the largest group north of the Volga and the basic dialect of the republic; Eastern Mari, used by a small group near Ufa, originally speakers of the Meadow dialect who emigrated in the late 18th century; and the Mountain dialect, to the west and on the south bank of the Volga. The Mountain and Meadow dialects both serve as literary languages and differ from each other only in minor details.

The Permic languages. Speakers of the three closely related Permic languages, Udmurt, Komi, and Permyak, number more than 900,000. Udmurt is concentrated largely in the vicinity of the lower Kama River just east of Mari El republic, in Udmurtia. Only very minor dialectal differences are found within Udmurt.

The Komi language area extends into the Nenets and Yamalo-Nenets autonomous *okrug*s (districts) far to the north. Lesser groups of Komi are found as far west as the Kola Peninsula and east of the Urals. Two major dialects are recognized, although the differences are not great: Komi (Zyryan), the largest group, which serves as the literary basis within Komi republic; and Komi-Yazva, spoken by a small, isolated group of Komi to the east of Komi-Permyak autonomous *okrug* and south of Komi republic. Permyak (also called Komi-Permyak) is spoken in Komi-Permyak, where it has literary status.

SAMOYEDIC

Nenets, with the largest number of speakers of all the Samoyed languages, has grown substantially in size over the past century, from some 9,200 speakers in 1897 to about 27,000 in 1989. Two distinct groups of Nenets differ in dialect as well as in cultural traditions: the Forest Nenets, a smaller, more concentrated group in the wooded area north of the central Ob River; and the Tundra Nenets, a group whose territory stretches roughly 1,000 miles eastward from the White Sea. These are the "Samoyadj" of Nestor's chronicles, but little is known of the history of any of the Samoyed peoples until recent centuries.

Nenets alone among the Samoyedic languages can claim a native literature, although both it and Selkup have been in written form since the 1930s. Evidence of the cultural prestige of certain Nenets tribes is seen in the adoption of a Samoyed language by Khanty speakers on the Yamal Peninsula. Enets is spoken by a dwindling group of fewer than a hundred Samoyeds near the mouth of the Yenisey River, just east of the Nenets. Nganasan, spoken by the northernmost Eurasian people, is found north and east of the Enets-speaking group, centring on the Taymyr Peninsula. The number of Nganasans has remained fairly constant, and they seem to have a high degree of ethnic identity (some 75 percent of 1,300 Nganasans still claimed Nganasan as their mother tongue in the late 1980s).

Native literature in Nenets

Selkup, the last of the southern Samoyed languages, is represented by scattered groups of speakers who live on the central West Siberian Plain between the Ob and the Yenisey. Only slightly more than one-third of Selkup speakers still considered the language their mother tongue in the late 1980s.

YUKAGHIR: A PROBABLE RELATIVE

The Yukaghir, in two small areas of Yakut-Sakha republic and Magadan *oblast* (province) of northeastern Siberia, have experienced a growth of total numbers during this century (still well under 1,000). But at the same time the number of speakers has declined by nearly 25 percent.

Linguistic characteristics

The linguistic structure of Proto-Uralic has been partially reconstructed by a comparison of the similarities and differences among the known Uralic tongues. Not all existing similarities can be attributed to a common Uralic origin; some may also reflect universal pressures and limitations on language structure (*e.g.*, the tendency to weaken stopped consonants between vowels, the modifying of a sound to become more similar to a preceding or following sound) or the influence of neighbouring, even genetically unrelated language structures (*e.g.*, the various types of vowel harmony [see below] in Finno-Ugric probably reflect such areal pressure).

PHONOLOGICAL CHARACTERISTICS

Finnish and Hungarian sound correspondences

The correspondences of sounds in cognate Uralic words are illustrated in Table 36. Thus, a *p* in the beginning of a Finnish word corresponds to *f* in Hungarian (*puu* : *fa*); a Finnish *k* is matched by Hungarian *h* before a back vowel (*a, o*), otherwise by *k;* within the word, Finnish *t* is matched by Hungarian *z,* and *nt* by *d;* Finnish initial *s* sometimes corresponds to Hungarian *sz* and sometimes to no consonant at all (*syli* : *öl*). In most of these instances, Finnish has retained the consonants of the Proto-Uralic consonant system. One exception is *nt,* which was originally **mt;* the *m* has become *n,* matching the position of articulation of the adjacent *t.* (An asterisk marks a form that is not found in any document or living dialect but is reconstructed as having once existed in an earlier stage of a language.) A second Finnish innovation is the loss of the distinction between the two original *s* sounds, **s* and **ś* (a palatalized *s,* as in *ship*). (Palatalization is the modification of a sound by simultaneous raising of the tongue to or toward the hard palate.) Hungarian maintains this distinction, but the original **s* words have lost this sound. By careful examination of such systematic relationships, it is possible to sketch out much of the phonological structure of early Uralic. The reconstructions in the last column of Table 36 are based on the view that the vowel system of Baltic-Finnic is relatively more conservative, whereas the consonant contrasts have been best preserved in Sami.

Consonants. The following consonant sounds are generally posited for the early stages of Uralic: **p, *t, *č* (pronounced as the *ch* in *chip*), **k, *s, *š* (pronounced as the *sh* in *ship*), **ð* (pronounced as the *th* in *then*), **l, *r, *m, *n, *ŋ* (pronounced *ng* as in *sing*), **j* (pronounced as the *y* in *yet*), **v,* and the palatalized alveolar sounds **t′, *ś, *ð′, *l′, *ń,* plus a few others less well established. Modern Finnish has a much smaller inventory of consonants, having lost the palatalized alveolar sounds and **č, *š, *ð,* and **ŋ.* Hungarian, on the other hand, has a larger number of consonants by virtue of a newly introduced distinction between sounds made with and without vibration of the vocal cords (voicing), such as voiceless *p, t, s* as opposed to voiced *b, d, z; e.g., dél* "noon" : *tél* "winter." Other Uralic languages, such as Komi, have also acquired a voicing contrast (*e.g., doj* "pain" : *toj* "louse"), but the geographic distribution of those languages in which the voicing contrast plays an active role leaves little doubt that it originated under the influence of Indo-European and Turkic languages.

Vowels. Essentially nothing is known of the Proto-Uralic vowels, and there is little agreement about the nature of the Proto-Finno-Ugric vowel system. It is clear, however, that, in contrast to a relatively limited number of consonants, Finno-Ugric must have had a fairly large number of vowels (nine to 11 are usually posited). One hypothesis is that the original vowel system was essentially like that of Finnish, which has eight vowel sounds: *i, ü, u, e, ö, o, ä, a* (*ü*—spelled *y* in the standard orthography—and *ö* are front rounded vowels, as in German; *ä* is a low front vowel, as *a* in *cat*). Hungarian has a similar system, although not all dialects have a separate *ä* sound, which is not distinguished from *e* in the orthography. A second approach posits a Proto-Uralic vowel structure closely resembling that of Khanty, with seven full vowels and three reduced vowels.

Early Finno-Ugric vowel length

The early Finno-Ugric system of vowels most likely possessed quantitative vowel contrasts (long versus short, or full versus reduced). Such contrasts are present in Baltic-Finnic, Sami, and Ugric and within Samoyedic—*e.g.,* Finnish *tulen* "of fire" and *tuulen* "of wind," *tuleen* "into fire," and *tuuleen* "into wind"; Hungarian *szel* "slice" and *szél* "wind," *szelet* "wind" (accusative case), and *szelét* "its wind" (accusative). The possibility of influence by neighbouring languages cannot be ruled out in the case of vowel length, because western Finno-Ugric languages have been in close contact with Slavic and Germanic languages with similar vowel contrasts, and the eastern languages form an areal group among themselves. The remaining languages lack vowel quantity and are in intimate contact with Russian, which has lost the original contrastive vowel quantity of Indo-European. The Izhma dialect of Komi, adjacent to Nenets, has superficial contrasts such as *pi* "son" versus *pī* "cloud," but this vowel length is the result of a change of an *l* at the end of the syllable to a vowel.

Stress. In numerous Uralic languages—including

Table 36: Representative Cognates in Selected Uralic Languages

English	Finnish	Estonian	Sami (Lapp)	Mari	Komi	Khanty	Hungarian	Nenets	Proto-Uralic
Head; end	*pää*	*pea*	—	—	*pom*	—	*fej*	*pä-*	**päŋe*
Tree	*puu*	*puu*	—	*pu*	*pu*	—	*fa*	*pā*	**puve*
Fish	*kala*	*kala*	*guolle*	*kol*	—	*kul*	*hal*	*xal′ä*	**kala*
House, hut	*kota*	*koda*	*goatte*	*kuðo*	*-ka*	*kat*	*ház*	—	**kota*
Who	*ken*	*ke(s)*	*gi*	*ke*	*kin*	—	*ki*	*xib′ä*	**ken*
Hand	*käte-*	*kät-*	*giettâ*	*kö*	*ki*	*köt*	*kéz*	—	**käte*
Louse	*täi*	*täi*	*dik′ke*	*ti*	*toj*	*tögtəm*	*tetű*	—	**täjka*
Know	*tunte-*	*tunde-*	*dow′dâ-*	—	*təd*	—	*tud*	*tumda-*	**tumte-*
Give	*anta-*	*anda-*	*vuow′de-*	*omta-*	*ud-*	*ŏntas*	*ad*	—	**amta-*
Eye	*silmä*	*silm*	*čal′bme*	*šinča*	*šin*	*sem*	*szem*	*sew*	**silmä*
Heart	*sydäm-*	*südam-*	*čâððam-*	*šüm*	*śələm*	*səm*	*szív*	*sēj*	**śüðam-*
Lap	*syli*	*süli*	*sâllâ*	*šəl*	*syl*	*jöl*	*öl*	—	**süle*
Vein	*suoni*	*soon*	*suodnâ*	*šön*	*sən*	*jan*	*ín*	*tēn-*	**sōne*
Mouse	*hiiri*	*hiir*	—	—	*šyr*	*junkər*	*egér*	—	**šiŋer*
Ice	*jää*	*jää*	*jiegŋâ*	*ij*	*ji*	*jöŋk*	*jég*	—	**jäŋe*
Blood	*veri*	*veri*	*vârrâ*	*vür*	—	—	*vér*	—	**vere*
Water	*vete-*	*vet-*	—	*vüt*	*va*	—	*víz*	*jīd-*	**vete*
Go	*men-*	*min-*	*mânnâ-*	*mija-*	*mun-*	*mən-*	*men-*	*min-*	**mene-*
One	*yhte-*	*üht-*	*ok′tâ*	*ikte*	*ət′ik*	*it*	*?egy*	—	**ükte*
Two	*kahte-*	*kaht-*	*guok′te*	*kok*	*kyk*	*kät*	*két*	—	**kakte*
Three	*kolme*	*kolm*	*gol′bmâ*	*kum*	*kujim*	*koləm*	*három*	—	**kolm-*
Four	*neljä*	*neli*	*njæl′lje*	*nyl*	*ńol*	*ńəlä*	*négy*	—	**neljä*
Five	*viite-*	*viit-*	*vit′tâ*	*vič*	*vit*	*vet*	*öt*	—	**vit(t)e*
Six	*kuute-*	*kuut-*	*gut′tâ*	*kut*	*kvajt*	*kut*	*hat*	—	**kut(t)e*

*Unattested, reconstructed form.

Finnish, Estonian, Hungarian, and Komi—stress is automatically on the first syllable of the word; it is likely that Proto-Uralic also had word-initial stress. Closely related to this initial stress is the apparent severe limitation on early Finno-Ugric noninitial vowels; the full range of contrasts was permitted only in the first syllable. In certain languages, such as Eastern Mari and the Yazva Komi dialect, stress is not bound to a given syllable, and determining the place of stress requires information concerning vowel quality as well—*e.g.*, Yazva *śibdinə̂* "to bind," *ľiččina* "to descend," *ľiśina* "wood" (the *i*'s, which receive stress, were long at an earlier period; *ś, ć, ľ* are palatalized consonants). Stress at the end of a word is also found—*e.g.*, in Eastern Mari and Udmurt. Nganasan has a mora-counting stress, falling on the third unit of vowel length from the end of the word (where short vowels count as one unit, long vowels as two).

Vowel harmony. Vowel harmony is among the more familiar traits of the modern Uralic languages. Although most Uralic scholars trace this feature back to Proto-Uralic, there is good reason to question this view. Vowel harmony is said to exist when certain vowels cannot occur with other specific vowels within some wider domain, generally within a word. For example, of the eight vowels of Finnish, within a simple word, any member of the set *ü, ö, ä* prohibits the use of any member of the set *u, o, a,* but *i* and *e* may occur with either set. That is, within a word, vowels that are either rounded (such as *ü, ö, u, o*) or low (such as *ä, a*) must agree with each other in frontness or backness. (The distinction is marked phonetically by putting two dots over the front vowels.) The unrounded front vowels, *i* and *e,* may occur with any of the other vowels. Thus, from *talo* "house" one may form *talossa* "in (the) house," but for *kynä* "pen" the comparable form is *kynässä* "in (the) pen"; similarly, *talossansa* "in his house" contrasts with *kynässänsä* "in his pen" and *talossansako* "in his house?" with *kynässänsäkö* "in his pen?", whereas *taloni* "my house" and *kynäni* "my pen" have the same ending because *i* can occur with either of the two sets of vowel classes. Hungarian has essentially the same system, differing only in certain minor details (short *e* is the front vowel counterpart of *a*)—*e.g., asztal* "table," *asztalok* "tables," *asztalokban* "in the tables," but *föld* "land," *földök* "lands," *földökben* "in the lands." Similar though less general front-back vowel-harmony systems are found in given dialects of Mordvin, Mari, Mansi, Khanty, and Kamas.

Vowel harmony versus assimilation of vowels

Frequently confused with the true harmony situations above are partial and total assimilations of vowels in adjacent syllables. These assimilations illustrate a universal tendency of vowel interaction and are of relatively recent origin; they are best held apart from the question of vowel harmony. Examples of vowel assimilations abound. In Finnish an unstressed *e* in the illative case ("place into") is totally assimilated to a preceding vowel, even across an intervening *h*: *talo* + *hen* becomes *taloon* "into the house," *talo* + *i* + *hen* yields *taloihin* "into the houses," *työ* + *hen* becomes *työhön* "into the work." The Hungarian allative case ("place to or toward which") shows an assimilation of the phonetic feature of lip rounding with front vowels in addition to the standard vowel harmony; thus, *ház-hoz* "to the house," *kéz-hez* "to the hand," *betű-höz* "to the letter." Apart from such nonharmony alternations, no support for rounding harmony is found in Uralic.

Considered from an areal viewpoint, two aspects of Uralic vowel harmony must be considered. First, those languages that show productive or active vowel harmony, with the exception of Baltic-Finnic, have had recent Turkic neighbours whose languages exhibited vowel harmony. For languages such as Mansi and Khanty, dialects with vowel harmony are located close to Tatar groups. Second, the original homeland of Uralic lies in the centre of an enormous hypothetical areal grouping, labeled by the Russian-American linguist Roman Jakobson as the "Eurasian language union." The languages of this "union" are said to be characterized by two features: (1) the absence of a tonal accent (changes in pitch that change meaning, as is found in Chinese, Swedish, or Serbian) and (2) the contrast of plain and palatalized consonants (as in Russian). The distinction between palatalized and nonpalatalized consonants has the same acoustic basis as the contrast of front and back vowels (*i.e.*, palatalized consonants and front vowels share a heightened tonal quality). Indeed, in Erzya Mordvin, vowel harmony and palatalization appear to be conditioned by essentially the same rules. Instead of seeking a genetic explanation of vowel harmony in Uralic, a somewhat more recent areal origin—in part under Turkic influence—must be considered. Of significance is the further consideration that, among the northwestern languages, far from Turkic influence, it is precisely Sami and the Baltic-Finnic Estonian and Livonian that do not have vowel harmony and that have developed special syllable-accent systems (thus, they lack both traits of the Eurasian union).

Consonant gradation. The alternation of consonants known as consonant gradation (or lenition) is sometimes thought to be of Uralic origin. In Baltic-Finnic, excluding Veps and Livonian, earlier intervocalic single stops were typically replaced by voiced and fricative consonantal variants, and geminate (double) stops were shortened to single stops just in case the preceding vowel was stressed and the following vowel was in a closed syllable; that is, **p* alternated with **v* and **b; *t* with **ð* and **d; *k* with **ɣ* and **g; *pp* with **p;* and so on. Finnish thus shows pairs such as *mato* "worm" and *madon* "of the worm," *matto* "rug" and *maton* "of the rug," *poika* "boy" and *pojan* "of the boy," *lintu* "bird" and *linnun* "of the bird," *selkä* "back" and *selän* "of the back." Estonian shows the same type of alternation, with considerable difference in detail—*e.g., sada* "hundred" and *saja* "of a hundred," *madu* "snake" and *mao* "of the snake," *lind* "bird" and *linnu* "of the bird," and *selg* "back" and *selja* "of the back." Most of the Sami languages exhibit similar alternations, but the process applies to all consonants and, moreover, works in reverse: single consonants are doubled in open syllables—*e.g., čuotte* "hundred" and *čuođe* "of a hundred," *borra* "eats" and *borâm* "I eat." The change of *t* to *đ*, however, is not a part of Sami gradation but rather a general process that voices and weakens all single stops between voiced sounds (in this case, vowels).

Variation in Sami (Lapp)

Despite their essential differences, the Baltic-Finnic and Sami gradations appear to be areally related. The Baltic-Finnic type, which represents a more plausible phonetic change, indicates that early Sami may have acquired its gradation under Baltic-Finnic influence. The evidence within Baltic-Finnic points to a relatively late, post-Proto-Baltic-Finnic origin. The existence of analogous consonant weakening in various Samoyedic languages (Nganasan, Selkup) is the result of independent innovation.

Syllable-accent structures. Closely related to the gradation phenomena is the development of syllable-accent structures in Estonian, Livonian, and Sami. Estonian is known for its unique quantity alternations of three contrastive vowel and consonant lengths—thus, *vara* "early" versus *vaara* "of the hillock" (*aa* = long *ā*) versus *vaara* "hillock (partitive)" (here *aa* = extra-long *â*); *lina* "linen" versus *linna* "of the city" (*nn* is pronounced as two short *n*'s) versus *linna* "into the city" (here *nn* is pronounced as long *n̄* plus short *n;* the contrast with the previous *nn* is not shown in the standard orthography). The extra-quantity contrast is in fact found with all stressed syllable types containing at least one vowel or consonant following its first vowel; thus, *taevas* "sky" (with short *e*) versus *taevas* "in the sky" (with long *ē*); *osta* "buy!" (with short *s*) versus *osta* "to buy" (with long *s̄*), whereas a two-syllable form such as *osa* "part" (*o/sa*) with only a single vowel in the first syllable is incapable of such a quantity contrast. A multitude of analyses of Estonian quantity have been proposed, although not all have recognized the phenomenon as a function of whole syllables bound to stress—in other words, that it is an accent phenomenon. One orthographic dictionary (by E. Muuk), for example, utilizes this principle, placing a grave accent mark before syllables with extra quantity. Otherwise, Estonian orthography marks the three degrees of duration only for stops: *b, d, g* indicate single short (voiceless lenis) stops (*tuba* "room"); *p, t, k* are plain geminates, or double consonants (*tupe* "of the sheath"); and *pp, tt, kk* mark extra-long geminates (*tuppa* "into the room," *tuppe* "into the sheath"). Because the extra quantity is in part tied to an original open next syllable, it frequently operates together with gradation alternations—*e.g.*, *linnu* "of the bird" versus *lindu* "bird (partitive)," with extra quantity.

The syllable quantity accent in Sami superficially resembles that in Estonian and, like the former, occurs only under stress and is in part conditioned by the openness of the next syllable. In North Sami (Utsjoki), alternations in paradigms involve three grades of quantity shaping: *mânâm* "I go" (*â* is a Sami letter for a somewhat rounded *a*) versus *mânna* "he goes" versus *mân'ne* "goer"; *dieđam* "I know" versus *dietta* "he knows" versus *diet'te* "knower"; *juol'ge* "leg" versus *juolge* "of the leg." This series of contrasts shows a three-stage decrease in initial-vowel duration and a three-stage increase in the duration of the first consonant after the first vowel or vowels. The other northern and eastern Sami languages display similar alternations, but there is considerable diversity in the phonetic details.

GRAMMATICAL CHARACTERISTICS

Early Uralic sentence structure

Word order. The grammatical structures of the various Uralic languages, despite numerous superficial differences, generally indicate a basic Early Uralic sentence structure of (subject) + (object) + main verb + (auxiliary verb)—the parenthesized elements are optional, and the last element is the finite (inflected) verb, which is suffixed to agree with the subject in person and number. This pattern has been best preserved in the more eastern languages, especially Samoyed, Yukaghir, and Ob-Ugric; *e.g.*, Nenets *ťiky pevśumd'o-m saravna t'eńe-va*ˀ "we well remember that evening" (literally, "that evening-[accusative] well remember-we"); Mari *joltaš-em-blak lum tol-mə-m buč-aš tüŋal-ət* "my friends begin to wait for the coming of snow" (literally, "friend-my-[plural] snow coming-[accusative] wait-to begin-they"); Yukaghir *met Tolstoj-wiejuol-kniglеŋ juonumeŋ* "I see a book written by Tolstoy" (literally, "I Tolstoy-written-book see-[present auxiliary]"). This order is common but optional in the languages of central Russia. Sami, Baltic-Finnic, and Hungarian now show the typical European subject-verb-object order: *e.g.*, Finnish *isä osti talo-n* "father bought a house(-genitive)," Hungarian *János keres egy ház-at* "John seeks a house(-accusative)." Although the latter languages have relatively "free" word order, the object precedes the verb only for special emphasis; *e.g.*, Hungarian *János egy házat keres* "John is looking for a *house* (and not something else)," Estonian *ma ta-lle leiba ei anna* "I won't give *him* any bread" (literally, "I him-to bread not give"). Estonian sentence structure somewhat resembles that of German, with its tendency to place the finite verb in second position while the rest of the verb complex remains at the end of the sentence; *e.g.*, *mehe-d ol-i-d ammu koju jõud-nud* "the men had got home long ago" (literally, "man-[plural] be-[past]-they long-ago home arrive-[past participle]").

The verb "be." In place of a verb "have," the Uralic languages use the verb "be," expressing the agent in an adverbial (locative or dative) case; *e.g.*, Finnish *isä-llä on talo* "father has a house" (literally, "father-at is house"), Hungarian *János-nak van egy ház-a* "John has a house" (literally, "John-to is one house-his"). In Proto-Uralic the copula verb "be" was lacking in simple predicate adjective or noun sentences, although the predicate was probably marked to agree with the subject. The following Hungarian sentences reflect this situation: *a ház fehér* "the house [is] white," *a ház-ak fehér-ek* "the houses [are] white." In Nenets and Mordvin such nonverbal predicates, even nouns, are conjugated for subject agreement and tense in the manner of intransitive verbs; *e.g.*, Nenets *mań xańenadm*ˀ "I am a hunter," *pydańi xańenadi*ˀ "you two are hunters," *mań xańenadamź* "I was a hunter," *pydara*ˀ *xańenadać* "you (plural) were hunters." Otherwise, a wide range of grammatical usage is found. In Baltic-Finnic and Sami the use of a copula verb is obligatory, in Permic it is optional, and in Hungarian the copula is absent only in the third person ("he, she") in a nonpast tense.

Negative sentences and questions. Negative sentences in Early Uralic were indicated by means of a marker known as an auxiliary of negation, which preceded the main verb and was marked with suffixes that agreed with the subject and perhaps tense. This is best reflected in the Finnic, Samoyedic, and Yukaghir languages; *e.g.*, Finnish *mene-n* "I go," *e-n mene* "I don't go," *mene-t* "you go," *e-t mene* "you don't go"; Yukaghir *met elūjeŋ* "I didn't go" (with negative prefix *el-* [*äl-* in Finnish]; compare *met merūjeŋ* "I went"). Ugric employs undeclined negative particles (*e.g.*, Hungarian *nem*), and in Estonian only negative imperative forms are still conjugated, although colloquial Estonian has initiated a tense distinction; *e.g.*, *ma/sa ei tule* "I/you don't come" and *ma/sa e-s tule* "I/you didn't come."

Interrogative pronouns

In Proto-Uralic, questions were formed with interrogative pronouns, beginning with **k-* and **m-*, illustrated by Finnish *kuka* "who," *mikä* "what" and Hungarian *ki* "who," *mi* "what." Yes–no questions were formed by attaching an interrogative particle to the verb, as in Finnish *mene-n-kö* "am I going?", *e-n-kö minä mene* "am I not going?" (in Finnish the verb also shifts to initial position). The use of intonation (changes in pitch) in interrogative sentences is currently widespread. In Hungarian it is the only way to form direct yes–no questions, although in indirect questions a particle *-e* is used; *e.g.*, *a házak fehérek?* (with sharply rising intonation of the next to the last syllable, dropping again on the final syllable) "are the houses white?", *nem tudom, fehérek-e a házak* "I don't know if the houses are white."

Conjunction. Conjunction, the connecting of clauses, phrases, or words, was formerly without the aid of specialized conjunctions. In the modern languages the conjunctions are largely borrowings from Germanic (Finnish *ja* "and") and Russian (Mari *da* "and; in order to," *a* "but," *ńi . . . ńi* "neither . . . nor," *jesle* "if"). Both coordination and subordination in sentences are marked by a wide range of constructions, especially by means of infinitive verbs, participles, and gerunds—*e.g.*, Mari *keče peš purgəžan poranan ulmaš* "the weather was very stormy and snowy" (literally, "weather very stormy snowy was"), *ača-ž aba-št* "their father and mother" (literally, "father-his mother-their"), *nuno batə-ž-den* "he and his wife" (literally, "they wife-his with"); Finnish *kirja-n lue-ttu-a-ni* "when I had read the book . . . " (literally, "book-[genitive] read-[past passive participle-partitive case]-my"), *luke-akse-ni kirja-n* "in order for me to read the book" (literally, "read-to-[translative case]-my book-[genitive]").

The case system. *Suffixes and postpositions.* Case suffixes and postpositions were and are used to show the function of words in a sentence. Prefixes and prepositions were unknown in Proto-Uralic. Adjectives, demonstrative pronouns, and numerals originally did not show agreement in case and number with the noun, as is still the case in Hungarian—*e.g.*, *a négy nagy ház-ban* "in the four large houses." Finnish, however, has initiated a case–number agreement system much like that in neighbouring Indo-European languages—*e.g.*, *neljä-ssä iso-ssa talo-ssa* "in the four large houses." The case system of the Proto-Uralic language contained an unmarked nominative case, an accusative, a case of separation (ablative), a locative (essive) case, and a case of direction (lative), plus possibly several others. The modern languages show a range from three cases in Khanty, six in Sami, 14 in Finnish, up to 16 to 21 for Hungarian (the case status of several suffixes is debatable). The average number of cases is about 12. For the most part, these cases are the same for all nouns, singular and plural, and many are similar in function to English prepositions. Nouns are not classified for gender, and third-person pronouns generally do not distinguish between "he" and "she."

The distinction between a case and a postposition is often based on arbitrary and superficial criteria. Postpositions, preposition-like elements following a noun, are more independent than cases, and they also function as adverbs. They often resemble inflected nouns (*e.g.*, Finnish *taka-* "behind": *talo-n taka-na* "house[-genitive] behind at," *talo-n taka-a* "house behind from," *taka-osa* "back part").

The original case relationships of essive–lative–ablative form a three-way set of contrasts that has been extended

Table 37: Case Endings in Several Uralic Languages

Finnish	Komi	Hungarian	Nenets	English translation
talo-ssa	*kerka-yn*	*ház-ban*	*xarda-xa-na*	"in (the) house"
talo-i-ssa	*kerka-jas-yn*	*ház-ak-ban*	*xarda-xa-ʔ-na*	"in (the) houses"
talo-sta	*kerka-yś*	*ház-ból*	*xarda-xa-d*	"from (the) house"
talo-i-sta	*kerka-jas-yś*	*ház-ak-ból*	*xarda-xa-t* (from *xa-ʔ-d*)	"from (the) houses"

Parallel series of cases

into several parallel series of cases in the modern languages. For example, Finnish uses essentially the original three in relatively abstract functions (essive, a state of being, *-na;* translative, a change of state, *-ksi;* partitive, a case of separation, [*-t*]*a*) and also adds an *-s-* element to indicate internal relationship (*-ssa* from **s* + *na* "in"; *-hen,* or a vowel + *n,* etc., from **s* + *ń* "into"; *-sta* "out of") and an *-l-* element to indicate external relationship (*-lla* from **l* + *na* "on, at," *-lle* from **l* + *k* "onto, to," *-lta* "off of, from"). Hungarian has nine cases similarly organized into three series of three, the internal set of which (*-ben* "in," *-be* "into," *-bõl* "out of") has recently developed from a noun with the meaning "intestines" (*bél*). In Finnish the personal pronouns are declined throughout on a pronoun stem—*e.g., minä* "I," *minu-ssa* "in me," *minu-n* "me (genitive)," and so on. In Hungarian, however, only the nominative and accusative forms are formed this way, and the remaining cases are formed by adding the possessive suffixes to a form of the case marker (sometimes expanded)—*e.g., te* "you (singular)," *teged-et* "you (accusative)," *benn-ed* "in you," *belé-d* "into you," *belõ l-ed* "out of you."

Noun inflection. The inflection of nouns for number (singular and plural) in the Uralic languages is much looser than in the Indo-European languages. Suffixes for the plural in the various Uralic languages are so diverse as to suggest that early stages of Uralic did not possess a specialized number marker—*e.g.,* Finnish *-t* and *-i-*, Mari *-blak,* Komi *-jas.* A dual-plural distinction ("two" as opposed to "more than two") is found in Sami, Ob-Ugric, and Samoyedic, but here again the specific elements cannot be traced to a common source. If Proto-Uralic had plural and dual suffixes, they were probably used only with the personal pronouns. In the modern languages personal pronouns often take a plural marker different from that of the nouns, and in Sami the dual formation is restricted to pronouns and personal affixes.

The category of definiteness (like English "the") is marked in numerous ways in the modern languages and originally appears to have been tied to the manner of number marking in Uralic (plural being reflected by indefiniteness). Hungarian alone has a definite article, *a*(*z*), a demonstrative in origin; Mordvin has three sets of inflectional endings: indefinite, definite singular, and definite plural (*kudo-so* "in a house," *kudo-so-ńt'* "in the house," *kudo-t'ńe-sə* "in the houses"). Nearly all the more eastern members have a definite marker that is identical with the third- or second-person possessive suffix (Komi *kerka-ys*/*yd* "the house" or "his/your house").

Possession. In possessive constructions the possessor noun precedes the possessed noun, or, in the case of a personal pronoun possessor, possessive suffixes are used—*e.g.,* Finnish *isä-n talo* "father's house" (*-n* = genitive), *talo-ni*/*si* "my/your house"; Hungarian *János ház-a* "John's house" (*-a* = possessive construction marker), *ház-am*/*ad* "my/your house." Although in earlier stages the possessive suffixes followed the case suffixes, more recent case formations (especially from original postpositions) have led to restructuring of this order—*e.g.,* Finnish *talo-i-ssa-ni* "in my houses," but Hungarian *ház-a-i-m-ban* "in my houses" (*-i-* = plural); Komi *kerka-yd-ly* "for your house" (*-yd-* = "your"), *kerka-ś-yd* "from your house," where two fixed orders coexist. The Proto-Uralic comparative construction was similar to the Finnish *talo-a iso-mpi* "house-from larg-er" (= "larger than a house"); compare Hungarian *egy ház-nál nagy-obb* "house-by larg-er" (in dialects also *ház-tól* "house-from"); Komi *kerka dor-yś yjyd-jyk* "house by-from larg-er." Parallel "than" type conjunctions are now common in the more western languages; *e.g.,* "larger than a house" in Finnish can also be expressed as *isompi kuin talo* (*kuin* = "than"), and in Hungarian *nagyobb mint egy ház* (*mint* = "than").

Compounding and derivation in noun formation

Compounding. The formation of nouns in Proto-Uralic included compounding (adding two or more words together) as well as derivation by the use of suffixes (word endings). In noun + noun constructions, including titles of address, the qualifying noun came first; compare Hungarian *házhely* "house site," *Szabó János úr* "Mr. John Szabó"; Finnish *taloryhmä* "group of houses," *Sirpa täti* "Aunt Sirpa." The rich system of derived words in Uralic together with the various inflectional suffixes led to relatively long words; compare Finnish *talo-ttom-uude-ssa-ni-kin* "even in my houselessness" (literally, "house-less-ness-in-my-even"), Hungarian *ház-atlan-ság-om-ban* "in my houselessness."

Verb inflections. The Proto-Uralic verb was inflected for tense-aspect (**-pa* indicated "nonpast," **-ka* indicated "perfect nonpast; imperative," **-ja* indicated "past") and mood (**-ne* indicated "conditional-potential"). The use of auxiliary verbs to indicate tenses was unknown, although Sami, Baltic-Finnic, and Hungarian now have essentially a Germanic-type tense system, with perfect formations based on the "be" verb; *e.g.,* Finnish *mene-n* "I go," *ole-n men-nyt* "I have gone" ("be-I go-[past participle]"), *men-i-n* "I went," *ol-i-n men-nyt* "I had gone," *men-isi-n* "I would go," *ol-isi-n men-nyt* "I would have gone." Under Germanic and Slavic influence both Estonian and Hungarian have developed separable verbal prefixes with adverbial and aspectual meanings; *e.g.,* Estonian *ära söö-* "eat (perfective)" and *ta sõ-i kala ära* "he ate the fish" versus *ta sõ-i kala* "he was eating fish," *ta hakkas kala ära söö-ma* "he began to eat (up) the fish"; Hungarian *meg-tanul* "learn" (perfective) and *János megtanul-t magyar-ul* "John learned Hungarian" versus *János tanult magyarul* "John was learning Hungarian," *János tanult meg angolul* "*John* learned English," *János nemetül tanult meg* "John learned *German*" (with special emphasis as indicated).

Proto-Uralic did not have specialized voice markers, such as the Indo-European passive; rather, the function of voice was interwoven with topicalization (a way of indicating the main subject of a sentence), emphasis, and definiteness of the subject and object as well as with verbal aspect. An indefinite subject of an intransitive verb or an indefinite object were marked with the ablative case (**-ta*), but a definite object took the accusative marker (**-m*) and other subject situations were unmarked (nominative). This system is best preserved in Finnish: *vesi* (nominative) *juoksee* "the water is running" versus *vettä juoksee* "there is water running," *juon vede-n* "I will drink the water" (*-n* is from older **-m*) versus *juon vettä* "I drink water." (Note that aspect as well as tense is affected by these case distinctions.)

Subjective and objective conjugations

The widespread use of separate subjective and objective conjugations among the Uralic languages (as in Mordvin, Ugric, and Samoyedic) are the result of an original system for singling out the subject or object for emphasis (focus), and not simply a device for object–verb agreement (similar to subject agreement). For example, Nenets *tymʔ xada-v* "*I* killed a deer (focus on the agent)" versus *tymʔ xada-dmʔ* "I killed a *deer* (focus on the object)," in which *-v* signifies "I . . . it" (the objective conjugation) and *-dmʔ* signifies "I" (the subjective conjugation). Note also the objective forms *xada-n* "I killed [them]," *xada-r* "you (singular) killed [it]," *xada-d* "you (singular) killed [them]," and so on for nine possible subjects (three persons times singular, dual, plural) times two object numbers (singular and nonsingular [not actually distinguished with third-person subjects]); and the subjective forms *xad-n* "you (singular) killed" and so on, for nine subject agreements. Yukaghir similarly employs distinct conjugations to reflect sentence focus; *e.g., met ai* "*I* shot (focus on subject)," *met meraiŋ* "I *shot* (focus on verb)," *met ileleŋ aimeŋ* "I shot *the deer* (focus on object)." Hungarian opposes definite and indefinite conjugations: two different sets of personal endings are used—one with transitive verbs with definite objects and the other elsewhere—*e.g., olvas-om*/*od a level-et* "I/you read the letter" versus *olvas-ok*/*ol egy level-et* "I/you read a letter." Along with its subjective and objective conjugations, Khanty has added a so-called passive conjugation (compare *kitta-j-m* "I am being sent,"

-*j*- = "passive") as an extension of the earlier focus-topicalization system. Mari and Komi have two past tense formations with related function. Again, the westernmost languages have passive constructions similar to those in both Slavic and Germanic.

Verbal derivation was richly developed already in Proto-Uralic with a wide variety of verbal nouns, infinitives, and participles. Each of the three tense-aspect markers was apparently used as a participial formative (compare Finnish *lähde* from **läkte-k* "source," *lähtijä* "one who leaves," *lähte-vä* from *-*pa* "leaving").

Several of the modern Uralic languages make extensive use of their native derivational processes to eliminate foreign loanwords; *e.g.*, for "telephone" Finnish has *puhelin,* which is derived from *puhel-* "talk," just as *soitin* "musical instrument" comes from *soitta-* "to play." The Uralic finite verb originally may have been based on participial constructions parallel to the noun-plus-predicate-adjective sentences (like Hungarian *a ház fehér* "the house [is] white"). Thus, one may reconstruct sentences like **ema tumte-pa* "mother [is] knowing," **ema tumte-pa-ta* "mothers [are] knowing" (with subject number expressed only in the predicate [agreement]) to explain the close similarity of participial and finite verb constructions such as Estonian *tundev ema* "knowing mother," *tundvad emad* "knowing mothers," *ema tunneb* "mother knows," *emad tunnevad* "mothers know."

WRITING SYSTEMS AND TEXTS

The earliest known manuscript in a Uralic language is a Hungarian funeral oration (*Halotti Beszéd*), which stems from the turn of the 13th century AD. A 12-word Karelian fragment also dates from the 13th century. Old Permic, the earliest attested form of Komi, received its own alphabet in the 14th century, and the first Finnish and Estonian texts are 16th-century printed works. Sami was first written in the 17th century.

Since the 17th century nearly all the more populous Uralic languages have a written form. All the above-mentioned languages and most semiautonomous groups in Russia have a native literature, the exception being Karelia, which uses Finnish. Currently, Uralic languages used within Russia are written with a modified Cyrillic alphabet; the others employ the Latin alphabet. (R.T.H.)

ALTAIC LANGUAGES

The term Altaic includes the languages of the Turkic, Mongolian, and Manchu-Tungus language families. Named for the Altai Mountains, the languages are spoken by more than 135,000,000 people spread over an area that extends from the northeastern region of the Asian continent through the northern and northwestern provinces of China, Mongolia, Central Asia, southern Siberia, the Volga region, and Turkey, down to the Middle East and the Balkans.

Question of genetic relationship

On the basis of correspondences in vocabulary and structural similarities, several scholars have concluded that the Altaic languages are genetically related and thus use the term Altaic as the name of a language family. Others consider the correspondences and similarities to be only traces of ancient contacts or areal (regional) convergences and reject the genetic relationship of these languages. For them, the term Altaic refers to a language group that displays a certain historical and typological unity. Still others admit the possibility that the languages are related but do not consider the relation demonstrable.

Attempts have been made to determine further genetic relations of the Altaic languages. The examination of structural and etymological correspondences gave rise to the hypothesis that the Uralic (including Finnish and Hungarian) and Altaic languages are related; this hypothetical group is called the Ural-Altaic language family. But the Ural-Altaic theory has an ever-decreasing number of supporters. Some adherents of the Altaic relationship are seriously considering the possibility of genetic ties to certain ancient strata of the Korean language.

As a result of the historically very active role of the Altaic peoples, their languages are found spread over a large geographic area. Altaic peoples overpowered their present territories in Asia and Europe in succeeding waves. At one time or another they have played a dominant role in the history of China, Iran, Byzantium, the Arab caliphate, and India, and their migrations had effects on the history of eastern Europe as well. Their present political formations in Asia and in the Middle East are of great importance.

Languages of the Altaic group. The accompanying tables give information about the literary languages and the major dialects of the three language families, their geographic division, and the number of speakers. Historical connections of the individual languages also are indicated. The names of subdivisions of the language families are of geographic or historical origin, in conformity with scientific traditions.

Because the name Turkish usually referred to a single language—Ottoman Turkish, the idiom spoken in Turkey, the Balkan countries, Cyprus, and elsewhere—the term Turkic is preferred in scientific literature for the whole language family.

On the evidence of certain archaic features in which Chuvash corresponds to the Mongolian languages, adherents of the theory of an Altaic relationship tend to look upon Chuvash as a separate unit among the Turkic and Mongolian languages.

Mutual intelligibility is possible within each of the Turkic language groups, but there are major difficulties in comprehension among the groups, and understanding is impossible among the speakers of the Yakut, Khalaj, and Chuvash tongues and those of the so-called common Turkic languages.

The name Mongolian is used for the members of the

Table 38: Turkic Languages*

language groups	approximate number of speakers
Common Turkic, or z-Turkic, languages	
Southeast (Uighur or Chagatai) group	
Uzbek (Uzbekistan 16,000,000, Afghanistan 1,500,000, Kyrgyzstan 570,000, Russia 100,000)	18,170,000
Uighur (China 7,600,000, Kazakhstan 180,000)	7,780,000
Southwest (Oğuz or Turkmen) group	
Turkish (Turkey 53,600,000, Germany 1,800,000, other western European states 700,000, Balkan states 1,010,000, Iran 430,000, Cyprus 150,000, Uzbekistan 110,000)	57,800,000
Azerbaijani (Iran 10,020,000, Azerbaijan 6,110,000, Iraq 340,000, Georgia 310,000, Russia 280,000, Armenia 90,000)	17,150,000
Turkmen (Turkmenistan 2,910,000, Iran 930,000, Afghanistan 330,000, Uzbekistan 120,000)	4,290,000
Gagauz (Moldova 140,000, Bulgaria 12,000)	152,000
Northwest (Kipchak) group	
Karaim (Lithuania, Ukraine)	fewer than 1,000
Kumyk (Russia)	189,000
Karachay-Balkar (Russia)	150,000
Tatar (Russia 4,760,000, Uzbekistan 410,000, Kazakhstan 230,000, Romania 25,000)	5,425,000
Bashkir (Russia)	990,000
Kazakh (Kazakhstan 6,660,000, China 1,170,000, Uzbekistan 850,000, Russia 560,000, Mongolia 139,000)	9,379,000
Karakalpak (Uzbekistan)	440,000
Kyrgyz (Kyrgyzstan 2,360,000, Uzbekistan 160,000, China 150,000)	2,670,000
Nogay (Russia)	55,000
Northeast (Siberian or Altai) group	
Khakass (Russia)	57,000
Altai (Russia)	45,000
Tuvinian (Russia 200,000, Mongolia 25,000)	225,000
Yakut (Russia)	360,000
Khalaj (Iran)	17,000
r-Turkic language	
Chuvash (Russia)	1,390,000
Total	126,735,000

*The geographic location of the language area and data specifying the distribution and approximate number of speakers are given in parentheses.

Table 39: Mongolian Languages*

language groups	approximate number of speakers
Western (Oyrat) group	
Kalmyk (Russia)	160,000
Oyrat (Mongolia 49,000, China 110,000)	159,000
Mogol (Afghanistan)	fewer than 1,000
Eastern group	
Mongol (China 5,060,000, Mongolia 1,858,000)	6,918,000
Buryat (Russia 360,000, Mongolia 40,000)	400,000
Daghur (China)	130,000
Monguor (or Tu; China)	200,000
Tung-hsiang (China)	390,000
Total	8,357,000

*The geographic location of the language area and data specifying the distribution and number of speakers are given in parentheses.

language family taken collectively, as well as for certain idioms spoken in Mongolia (Outer Mongolia) and China (Inner Mongolia). To make an exact distinction between the family and these languages, the latter are sometimes designated by the name Mongol. The reverse use of these terms, however, is also known.

Kalmyk, Mongol, and Buryat (Buriat) are contemporary literary Mongolian languages. There are no great differences among these languages, and mutual intelligibility thus exists, though the possibility of mutual comprehension is decreasing among speakers of isolated and nonwritten dialects.

The census datum for Manchu has an ethnic rather than a linguistic value. It is generally assumed that Manchu must be considered as the language of a minor group, and it is, in fact, almost an extinct language.

There are considerable differences between the northern and southern groups of this family and among the individual languages and dialects of the southern group. Despite the immense geographic area over which they are spread, languages and dialects belonging to the northern group are less differentiated.

LINGUISTIC CHARACTERISTICS

Phonology. Simplicity is the characteristic feature of the phonological (sound) systems of the Altaic languages. Richness in vowels is compensated for by the scantiness of consonants. In addition, clusters of consonants are rare.

Sound harmony

Sound harmony is one of the most characteristic features of the Altaic languages. Its essence is that in an individual word only certain combinations of vowels may occur, and, in the case of certain consonants, specific variants must be selected according to the type of vowels in the word. The best known forms of vowel harmony are palatal and labial harmony; *i.e.,* only back (velar) or front (palatal) vowels and rounded (labial) or unrounded (nonlabial) vowels may occur in a word. For example, in Turkish, back vowels include *a, ı, o, u*; front vowels include *e, i, ö, ü;* rounded vowels are represented by *ö, ü, o, u* and unrounded vowels by *a, ı, e, i.*

This harmony extends to the suffixes, too, producing variants when added to words belonging to different vowel classes. The change in the vowels of the last two syllables (both suffixes) in the following Turkish words is notable: *at-lar-da* "on the horses," *it-ler-de* "on the dogs," *gör-ür-üm* "I see," *getir-ir-im* "I bring."

Sound harmony in the Altaic languages is the product of a complicated historical development. The form of vowel harmony varies from one language to another. Sound harmony is less developed in the Manchu-Tungus group, in which it may occasionally reflect an ancient state or special aberration. Adherents of the Altaic relationship maintain that the ancient state of these languages has been characterized by a dynamic stress on the first syllable, while musical tone on the end syllable was the result of a later development. The ancient state seems to be reflected in the contemporary Mongolian tongues and a group of Manchu-Tungus languages; in Turkic and in another group of the Manchu-Tungus languages, stress shifted to the end syllable.

Table 40: Manchu-Tungus Languages*

language groups	approximate number of speakers
Southern (Manchu) group	
Manchu (China)	fewer than 1,000
Nanai (Russia 5,000, China fewer than 1,000)	6,000
Other dialects (Russia 12,000, China 33,000)	45,000
Northern (Tungus) group	
Evenk (China 30,000, Russia 9,000, Mongolia 2,000)	41,000
Even, or Lamut (Russia)	8,000
Total	101,000

*The geographic location of the language area and data specifying the distribution and number of speakers are given in parentheses.

Grammar. From the point of view of grammatical features, there is a great difference between the Altaic languages and the Indo-European languages (which include English, French, and German). Altaic languages are agglutinative; that is, word formation consists of adding suffixes to the word roots.

The root may take a large number of suffixes (*e.g.,* Turkish *gel-me-dik-ler-i-nden* "because of their not coming"), though the order of the combination of suffixes is governed by strict rules. While certain suffixes in the Indo-European languages may perform several functions (*e.g.,* Latin *casarum* "of the houses," in which the single suffix *-arum* indicates a feminine, plural, genitive form), in the Altaic languages these functions are distributed among separate suffixes (*e.g.,* Turkish *ev-ler-in* "of the houses," which consists of a stem plus a plural suffix plus a genitive suffix).

Altaic languages lack definite articles (such as English "the") and grammatical gender (*e.g.,* German masculine, feminine, and neuter words) and use the singular form of nouns with numerals (*e.g.,* "two girl"). In the Manchu-Tungus languages, a distinction is made in the 1st person plural of the personal pronoun between exclusive ("we"—without you) and inclusive ("we"—with you). The same peculiarity is found in Mongolian languages and, as a result of Mongolian influence, in the Turkic languages of Siberia.

There are no prepositions in the Altaic languages, but postpositions are used very frequently. For example, an Altaic speaker does not say the equivalent of "until night" or "for John" but rather "night until" and "John for." In the Tungus languages, nominal (noun) inflection has approximately 20 cases, while the Manchu languages show a simpler state, with about five to eight cases. Turko-Mongolian languages also show a simple case system: they have about 6 to 10 cases. Because the plural is expressed by independent suffixes, the case suffixes are identical in singular and plural.

Parts of speech in the Altaic languages

Parts of speech in the Altaic languages are less differentiated than they are in the Indo-European languages; essentially, the only separate parts of speech that can be distinguished are nominals (words used as nouns), verbs, and particles (words that are neither nouns nor verbs). In contrast with the Tungusic languages, the Turkic and Mongolian languages do not possess any morphological means, such as specific suffixes, to distinguish between nouns and adjectives.

Altaic languages are rich in verbal nominals (nouns, adjectives, and adverbs derived from verbs) as well as in nouns of action, participles, and gerunds; these perform important grammatical functions in Altaic sentences. A conspicuous ancient Altaic peculiarity is that verbal predicative expressions (finite verbs) are of verbal-nominal origin; that is, these verbal formations are in most cases verbal nouns in predicate positions and express their personal relations with added personal or possessive pronouns: *e.g.,* Old Turkic *bil-t-im* "I knew" is literally "having known my," and Old Turkic *bil-ir-biz* "we know" is literally "knowing we." In Altaic languages there is no word that corresponds to the English "to have." To express this concept, a periphrastic construction (circumlocution) is used; *e.g.,* an Altaic speaker might say "The book is with me" rather than "I *have* the book."

Despite major differences in syntax, there are also significant common Altaic features. The modifier (attribute, appositive), for example, always precedes the modified word (*e.g.*, a noun), and the noun in the genitive case (the possessive) precedes the possessor (*e.g.*, "captain of the ship" is "of the ship, captain"). In the Tungusic languages the attribute agrees in number and case with the modified noun. This feature is completely lacking in the Turkic languages, though agreement in number existed in the early periods of Mongolian.

In Altaic sentence structure, subordinate clauses that begin with conjunctions or relative pronouns ("The man *who was bald*") are alien; wherever they appear, they indicate foreign influence. Instead of subordination, coordination by verbal nominals (nouns) serves as a basis for sentence construction. The following Turkish sentences illustrate this feature. *Ahmedin bize gelmesini isterim,* translated as "I want Ahmed to come to us," is literally "I want Ahmed's coming to us"; *Şehre giden otobüs gecikti* "The bus, which goes to the town, was late" is literally "The bus going to the town was late"; *Eve varınca hemen telefon etti* "As soon as he arrived home he telephoned" is literally "Arriving home, he immediately telephoned." As a result of a more varied system of morphology, the Manchu-Tungus languages are characterized by a higher degree of syntactic freedom, while the Turkic and Mongolian languages have a higher degree of syntactic constraints (requirements for the positions of the elements of a sentence) and a simpler system of morphology.

Altaic languages have changed at a slow rate. The archaic character of the contemporary language is particularly conspicuous in the Manchu-Tungus and Mongolian languages. Because of this difference in historical character, one language may provide useful information for the history of another in regard to the development of earlier, supposedly common features.

Vocabulary. During an immense span of space and time, the vocabulary of the Altaic languages has shown a high degree of receptivity. Opponents of the genetic Altaic relationship look upon the ancient common elements of the vocabulary of these languages as loanwords resulting from once-frequent contacts between Turkic and Mongolian languages and between Turko-Mongolian and Manchu-Tungus.

These correspondences in vocabulary show closer ties between the Turkic and Mongolian languages than between the Turko-Mongolian and Manchu-Tungus languages.

Oldest loanwords of the Altaic vocabulary

The oldest loanwords of the vocabulary are connected with the Semitic, Indo-European, and Uralic languages. In the eastern part of the Altaic language area, many loanwords came from the Chinese to the Old Turkic and Middle Mongolian languages, as well as to contemporary Turkic and Mongolian, and also to the Manchu-Tungus languages, particularly to Manchu.

Through the medium of religious texts translated from Sanskrit, Tocharian, and Sogdian, many loanwords found their way to Uighur and from there to the Mongol language. Mongolian also had contacts with the Tibetan language through the Buddhist literature and through certain dialects. A strong linguistic interaction may be observed among the local Turkic and Mongolian languages in Mongolia and southern Siberia and among the Turkic and Iranian languages in Central Asia.

Other influences were felt as well. A number of Mongolian words were adopted into the Yakut language, which in turn had contact with the Tungusic languages. The spread of the Islāmic religion paved the way for Arabic and Persian loanwords to enter the Turkic languages. Because of interactions between the Ottoman Empire and the western European civilizations, a considerable number of French, Italian, German, and English words were borrowed into Turkish. Russian provided the major part of the vocabulary of modern civilization for the Turkic languages of Russia, although there was also considerable influence from other spheres. The Balkan, Anatolian, Caucasian, and certain Uralic languages had only minor influence on the Turkic languages.

In certain Turkic languages, the influence of linguistic contacts goes far beyond the simple borrowing of words. Although in the East the development of the Buddhist translation literature was followed by a comparatively modest linguistic influence by the source languages, the Arabic and Persian influence that followed the adoption of Islāmic religion and culture proved very strong. Arabic and Persian words flooded the vocabulary of the literary languages of the settled Islāmic-Turkic peoples and also filtered into the dialects.

This dense layer of loanwords also eventually led to changes in the structure (*i.e.*, the phonological system and sound harmony) of certain languages. Iranian influence gave rise to similar profound changes in Uzbek. During the Soviet era Russian influence was greatly intensified by the continuous contact taking place within an identical political and cultural framework and also by the increasingly widespread bilingualism of the Turkic peoples. This process is reflected in the growing number of calques (loan translation words) as well as in syntactic (sentence) structure.

In the course of linguistic contacts, the Altaic languages naturally also exerted a considerable influence on other languages—*e.g.*, Persian, Russian, Finno-Ugric, and the Balkan languages.

Earliest Turkic records

Writing systems. The Altaic peoples used a variety of writing systems. The first-known writing of the Turkic peoples is the runic script (named after the Germanic runes), which is most probably of Semitic—in particular, Aramaic—origin. The first dated texts originate from northern Mongolia in the 8th century; they are known as the Orhon inscriptions. Archaeological investigations indicate that this early script was spread over an extremely large area.

The Turkic peoples (mainly Uighur) who settled in the Tarim basin used Brāhmī and Manichaean scripts in the 8th to 10th century, as well as the Uighur writing, as it developed from the Sogdian cursive writing, from the 8th to 17th century. Nestorian (Christian) Turkic peoples in Central Asia used the Estrangelo script of Syriac origin in the 13th and 14th centuries.

After the adoption of Islām in the 11th century, the Arabic script became the most widespread writing system among the Turkic peoples. In Anatolia certain linguistic groups or minorities used the Greek alphabet (16th–20th centuries) and the Armenian script (17th–20th centuries). The latter also was employed by the Turkified Armenians in Ukraine in the 16th and 17th centuries.

Between the 16th and 19th centuries, Karaites used the Hebrew script; Latin and Cyrillic writing was introduced to them in the 19th century. In 1929 the Arabic script was replaced in Turkey by Latin writing. Among the Turkic peoples of Central Asia, the Latin script was introduced in the 1920s. In the 1930s it was replaced by the Cyrillic script, though the future of Cyrillic script for some of these languages was uncertain in the post-Soviet era. Turks living in China, Iran, and the Arab countries still use the Arabic script.

Mongolian script developed from Uighur writing in the 12th century. After some Latin antecedents, it was replaced by the Cyrillic script among the Buryat in 1938 and among the Mongols in 1944. Mongols living in China still use the Mongolian script. The 'Phags-pa or Pa-sse-pa alphabet, introduced on the state's initiative in 1269–72, was based on the Tibetan script and did not take root. It became extinct in a few decades. From the 17th century, the Oyrat used a reformed version of the Mongolian writing, called Oyrat script. After 1917 this writing system was replaced among the Kalmyk by the Cyrillic alphabet; the Latin alphabet was introduced in 1931, and in 1937 the Cyrillic script was reintroduced. Oyrat living in China still use the Oyrat script.

The Manchu script, which is a reformed or adapted version of Mongolian writing, came into existence in the 17th century; it is probably still used in private life in some parts of China. For Manchu-Tungus people living in what was then the Soviet Union, the Cyrillic script was introduced in 1937–38 after some experiment with the Latin alphabet.

The writings of the Mongolian-speaking Khitan, who ruled in China from the 10th to the 12th century, and those of the Juchen, who ruled in China in the 12th and 13th centuries and who spoke a Manchu-Tungus language, have not yet been definitely deciphered.

HISTORY

Original habitat of the Altaic peoples

The original habitat of the Altaic peoples is supposed to have been on the steppe area bordering on the Altai and Ch'ing-hai mountains between Tibet and China; in the north, it probably stretched to the Siberian taiga region adjacent to the steppe. It is assumed that Tungus peoples lived in the northern and northeastern parts of this area and the Mongols in the central and southeastern parts. Turkic peoples probably inhabited the northwestern and western parts, while the southern and southwestern region was occupied by Hunnic groups. Although the Hun language is not known, Huns are supposed to have spoken an Altaic idiom. In historical times, the Uralic, Indo-Iranian, Sino-Tibetan, and Korean languages have bordered on this Altaic language area.

Advocates of a genetic relationship between the Altaic languages believe that it was probably in the area delineated above that the Proto-Altaic language (*i.e.*, the parent language of the modern Altaic tongues) began to divide into the currently known main groups. Those who oppose the Altaic theory maintain that this area was the scene of contacts among the individual language groups. In either case, from this original habitat, the Altaic peoples moved to their present territories. In the steppe migrations, which were natural consequences of a nomadic way of life and which dramatically changed the ethnic and linguistic map of central Eurasia, the Turkic and Mongolian peoples played the leading role. The Manchu-Tungus peoples' separate migration toward the northeastern Asian taiga and tundra might have taken place in an early period. The south- and southwest-bound migrations of the southern Manchu-Tungus peoples were restricted to the East Asian regions.

The Turkic languages. *Division into two linguistic groups.* It was probably at a very early date that the Turkic languages split into the *z* and *š* group (commonly called *z*-Turkic) and *r* and *l* group (commonly called *r*-Turkic); these groups are so known because of certain regular phonetic contrasts involving those sounds, *e.g.*, *z*-Turkic *yüz,* "hundred," contrasts with *r*-Turkic *śĕr* "hundred"; *š*-Turkic *yaš,* "age," contrasts with *l*-Turkic *śul* "age." The first traces of the *r* and *l* division appeared on the south Russian steppes. It is assumed that the Huns also were speakers of an *r*- and *l*-type Turkic language and that their migration was responsible for the appearance of this language in the West. The *r*- and *l*-type language is now documented only by Chuvash, a language considered as a descendant of a Volga-Bulgarian language. The rest of the Turkic languages are of the *z*- and *š*-type.

The split of the *z*- and *š*-type Turkic languages, which appeared first in Inner Asia and later developed in the presently known language groups and languages, evolved over several centuries.

Literary languages and linguistic records. As was stated above, the earliest known records of the Turkic languages are runic inscriptions left by the Turks in Mongolia. The Uighurs overthrew the Turkish Empire there, and, after the disintegration of their empire in 840, they settled down in the Tarim Basin, where exposure to the Manichaean and Buddhist religions gave rise to a sizable literature, written in various scripts, that flourished from the 9th to the 14th century. In the 11th century, during the rule of the Qarakhanids (the first Turkic representatives of Islām), a new literature came into existence in eastern Turkistan. These literary efforts clearly show the intensification of Arabic and Persian influence, which culminated in the Khwārezmian literature (13th–14th centuries) and the Chagatai literature (15th–16th centuries), as well as in the postclassical products of Chagatai (17th–19th centuries). (Khwārezmian and Chagatai are Middle Turkic languages that possess significant literatures.) Despite the influence exerted by other Turkic language groups, these Inner and Central Asian literary languages exhibit an organic continuity; they might be considered as the antecedents of the contemporary Uzbek and New Uighur literary languages.

Arabic and Persian influence on Turkic literature

The main movement of the Oğuz branch of Turkic peoples from western Turkistan was toward the Caucasus, Iran, Anatolia, and the Balkans. The first continuous records of this language group come from 13th-century Anatolia. Its linguistic material is linked with the contemporary literary language of Turkey (*i.e.*, Turkish). The Azerbaijani and Turkmen languages also have long literary traditions. The first records of the Azerbaijani language date from the 14th century; the Turkmen language came into use only gradually, because the Chagatai language was commonly used instead by writers.

The migration of the Kipchak Turks was toward the south Russian steppes. A 14th-century textbook known as the *Codex Cumanicus,* used by missionaries, constitutes the first record of the languages of these peoples. Linguistic records (*e.g.*, dictionaries, grammars, literary works) of the Kipchaks, who as mercenaries went to Egypt and Syria where they also came into power, date from the 13th to 17th century. Armenian Turks of the Ukraine left remarkable written material from the 16th and 17th centuries. All these Kipchak linguistic records, although they represent languages that have disappeared, are still connected indirectly with the contemporary Kipchak literary languages. Because of the cultural conditions of these peoples, the first records of most of these languages originated in the 20th century.

The remainder of the contemporary Turkic literary languages also developed in the 20th century.

The diffusion of the Turkic peoples over the centuries and the changes in their political and cultural centres were accompanied by the emergence of a number of literary languages. These literary languages are not necessarily the direct antecedents of the contemporary—and recently developed—literary languages. In most cases, only their membership in the same language group (Uighur, Kipchak, Oğuz) may be established, thus indicating an indirect historical connection.

Stages of the languages. On the basis of linguistic records, Turkic languages that were spoken from the 8th to the 10th century are termed Old Turkic, those of the 11th to 15th century are called Middle Turkic, and those of the 16th to 20th century are known as New Turkic languages. To specify the form of the Turkic languages that resulted from language reforms that were instituted in the mid-20th century, it is customary to use the term newest, or modern. The division into periods is based on external facts of these languages rather than the internal history and thus merely gives a practical chronological overview.

The Mongolian languages. Although the Mongolian script probably evolved as early as the 12th century, its earliest known record, the inscription of Yesunke, dates from 1225. The *Secret History of the Mongols,* which was written in all probability in 1240, has been handed down in Chinese transcription. As regards the development of spoken Mongolian, the history of Mongolian is divided into three periods: Old Mongolian (to the 12th century), Middle Mongolian (13th–16th centuries), and New Mongolian (17th–20th centuries). Written Mongolian, which was originally based on a Middle Mongolian dialect, was characterized for centuries by an increasing conservativism in regard to its relation to the spoken language. Its history is divided into such periods as preclassical (13th–16th centuries), classical (17th–20th centuries), and modern (20th century). Classical Mongolian is continued by the contemporary literary Mongol, which is based on the Khalkha Mongol dialect, and by the written language of Mongols living in China.

Stages of Mongolian

The Oyrats created a separate script and literary language in the 17th century. Although this separate language gradually decreased in importance among the eastern Oyrats from the second half of the 18th century, it survived among the western Oyrats (the Kalmyks), and its course of development flowed into the contemporary literary language.

It was not until 1931 that the Buryats began to develop their own literary language. Until then, literate groups had used literary Mongolian.

The Manchu-Tungus languages. Within the Manchu-Tungus language family, only the history of the Manchu language is documented by linguistic records. As a result of the Manchu rule in China (1644–1911/12), Manchu rose to the rank of official language in the country. The earliest Manchu records date from the mid-17th century. Manchu acted as the lingua franca in China's communications with other countries. The downfall of the Manchu rule in China and the gradual Sinicization of the Manchu population ended the influence of the Manchu language. Other literary languages of the Manchu-Tungus language family came to be in the 1930s.

As has been shown, the Altaic languages, unlike the Indo-European languages, have relatively very recent linguistic records—Turkic dating from the 8th century, Mongolian from the 13th century, and Manchu-Tungus from the 17th century. This situation puts limits on what can be learned from comparative historical investigations, making it extremely difficult to settle the problem of the Altaic relationship.

Altaic languages in the 20th century. The development of literary languages became a central problem in the independent political entities of the Altaic peoples that arose after World War I. The languages of the Altaic peoples in the U.S.S.R. underwent a veritable revolution in the 1920s and 1930s. Writing reforms were soon followed by the strengthening of literary languages on the basis of the individual national languages. For several illiterate peoples, new literary languages had to be created for the first time. In Turkey, written and spoken language reforms included, above all, a renewal of the vocabulary; this formed a significant objective of the Turkish president Kemal Atatürk's social program. In the ongoing process of modernization, a new literary language has come into existence. Considerable reform of writing and language also took place in Mongolia.

Creation of new literary languages

The linguistic policy pursued in the mid-20th century had an immense impact on the history of the Altaic languages. These efforts not only saved minor or nonwritten languages from extinction but strengthened the positions of major Altaic languages among the languages of modern civilizations. The linguistic consolidation has considerably contributed to enhancing the political and cultural significance of these peoples.

For information on the literature of the peoples who speak (or have spoken) Altaic languages, see CENTRAL ASIAN ARTS; ISLĀMIC ARTS.

(G.Ha.)

DRAVIDIAN LANGUAGES

The Dravidian language family, as known to date, consists of 23 languages spoken by more than 165,000,000 people in South Asia. In terms of population figures the major languages of the family may be listed in the following order: Telugu, 52,986,000; Tamil, 44,400,000; Kannada (Kannaḍa), also called Kanarese, 27,900,000; Malayalam (Malayālam), 27,500,000; Gondi, 2,460,100; Tulu (Tuḷu), 1,427,000; and Kurukh (Kuruḵẖ), 1,358,000. The Dravidian languages are spoken in the Republic of India (mainly in its southern, eastern, and central parts), in Sri Lanka (Ceylon), and by settlers in areas of Southeastern Asia, southern and eastern Africa, and elsewhere. Brahui (Brāhuī), with 750,000 speakers in Pakistan, is isolated from all of the other members of the family. The four major languages—Telugu, Tamil, Kannada, and Malayalam—possess independent scripts and literary histories dating from the pre-Christian Era. Now recognized by the constitution of India, they form the basis of the linguistic states of Andhra Pradesh (established as the first Indian linguistic state in 1953), Tamil Nadu, Karnataka (formerly Mysore), and Kerala.

Official use of Dravidian languages

Of the Dravidian languages, Tamil has the greatest geographical extension and the richest and most ancient literature, which is paralleled in India only by that of Sanskrit. Its phonological and grammatical systems correspond in many points to the ancestral parent language, called Proto-Dravidian.

Nothing definite is known about the origin of the Dravidian family. There are vague indigenous traditions about an ancient migration from the south, from a submerged continent in what is now the Indian Ocean. According to some scholars, Dravidian languages are indigenous to India. In recent years, a hypothesis has been gaining ground that posits a movement of Dravidian speakers from the northwest to the south and east of the Indian Peninsula, a movement originating possibly from as far away as Central Asia. Another theory connects the Dravidian speakers with the peoples of the Indus Valley civilization. The Dravidian languages have remained an isolated family to the present day and have defied all of the attempts to show a connection with the Indo-European tongues, Mitanni, Basque, Sumerian, or Korean. The most promising and plausible hypothesis is that of a linguistic relationship with the Uralic (Hungarian, Finnish) and Altaic (Turkish, Mongol) language groups.

As an independent family, the Dravidian languages were first recognized in 1816 by Francis W. Ellis, a British civil servant. The actual term Dravidian was first employed by Robert A. Caldwell, who introduced the Sanskrit word *drāvida* (which, in a 7th-century text, obviously meant Tamil) into his epoch-making *A Comparative Grammar of the Dravidian or South Indian Family of Languages* (1856).

Languages of the family. Tamil is spoken by 39,400,000 people (1981 est.) in the Indian state of Tamil Nadu, by another 2,697,000 in Sri Lanka (Ceylon), by smaller numbers of people in Burma, Malaysia, Indonesia, and Vietnam (about 1,400,000), in East and South Africa (almost 250,000), and by still smaller numbers in Guyana and on the islands of Fiji, Mauritius, Réunion, Madagascar, Trinidad, and Martinique. The earliest literary monuments of the language belong roughly to the 3rd and 2nd centuries BC. There exist a number of local dialects, the major dialect regions being the northern and eastern areas combined, the western area, the southern area (split into at least four major dialects of Madurai, Tirunelveli, Nanjiland, and Ramnad), and Sri Lanka (Ceylon). Correlated with the social position of the speaker are a number of speech forms; a major division occurs between the Brahmin and the non-Brahmin varieties. In addition, there is a sharp dichotomy between the formal language and informal speech.

Dialects and forms of Tamil

Malayalam, which is closely related to Tamil, is spoken in the Indian state of Kerala by some 21,700,000 people. Possessing an independent written script, it also has a rich modern literature. There are at least three main regional dialects (North, Central, South) of Malayalam and a number of communal dialects.

In the Nīlgiris and adjacent regions, several minor tribes speak the following languages: Kota (1,400), Toda (1,145), Badaga (128,500), Irula (Iruḷa) (6,176). The less well-known languages of a number of other tribes may yet be established as independent members of the Dravidian family (*e.g.*, Kurumba, Paṇiya).

Kodagu (Koḍagu), a non-literary language of a mountainous region called Coorg, has 119,000 speakers.

Kannada (Kanarese), which is spoken by 25,700,000 people in the Indian state of Karnataka, exhibits a dichotomy between educated speech and colloquial Kannada; in the latter at least three social dialects are recognizable that may be characterized as Brahmin, non-Brahmin, and Harijan ("untouchable"). A number of regional dialects (among them are Dharwar, Bangalore, and Mangalore) also exist. Kannada has an orthography of its own and an important ancient and modern literature.

To the south of the Kannada territory, more than 1,400,000 people speak Tulu (Tuḷu), a South Dravidian language having no developed written literature.

Telugu (spoken by 52,986,000 people), the official language of the state of Andhra Pradesh, exhibits a dichotomy between the written and the spoken styles, in addition to a number of sharply distinct local and regional dialects (including Telangana, coastal area, Rayalaseema, and a "transitional" zone) and divisions between Brahmin, non-Brahmin, and Harijan speech. The language has its own

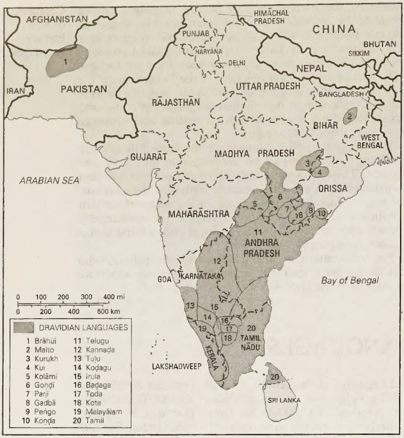


Figure 19: Distribution of the Dravidian languages.

Adapted from Ramanujan and Masica, "Toward a Phonological Typology of the Indian Linguistic Area," *Current Trends in Linguistics*, vol. 5 (1969); Mouton & Co., Publishers, The Hague

script, closely akin to that of Kannaḍa, and an important literary tradition.

In extreme northern Andhra Pradesh and in Mahārāshtra, the Kolāmī language is spoken by approximately 84,000 individuals. Parjī is spoken by about 36,000 individuals in Andhra Pradesh and Orissa. The Konda Dora, a scheduled tribe of some 23,000, live mostly in Andhra Pradesh and speak Koṇḍa. The Gadbā, who live mainly in Andhra Pradesh, number approximately 28,000. Peṅgo is spoken by fewer than 2,000 individuals living in Orissa, and Kui and Kuvi are spoken by a number of tribes in Andhra Pradesh and Orissa.

In Madhya Pradesh and parts of Andhra Pradesh, Mahārāshtra, and Orissa, many groups of Gonds (including about 2,620,000 persons) speak a number of Goṇḍī dialects. To the north, in Assam, Bihār, Madhya Pradesh, Tripura, and West Bengal, the Oraon tribe speaks Kurukh (1,700,000), and, near the borders of Bihār and West Bengal, 100,000 tribals speak Malto.

The only Dravidian language that is spoken entirely outside India is Brāhūī, with about 1,580,000 speakers

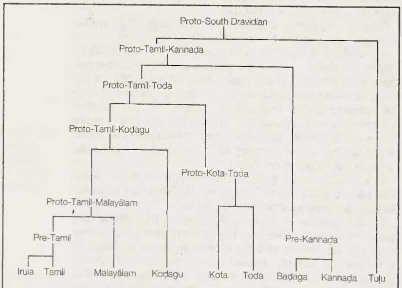


Figure 20: South Dravidian subfamily.

who live in Sindh and Balochistān provinces of southern Pakistan.

Historical survey of the Dravidian languages. Although in modern times speakers of the various Dravidian languages have mainly occupied the southern portion of India, while those of the Indo-Aryan (Indic) tongues have predominated in northern India, nothing definite is known about the ancient domain of the Dravidian parent speech. It is, however, a well-established and well-supported hypothesis that Dravidian speakers must have been widespread throughout India, including the northwest region. This is clear because a number of features of the Dravidian languages appear in the Rigveda, the earliest known Indo-Aryan literary work, thus showing that the Dravidian languages must have been present in the area of the Indo-Aryan ones. The Indo-Aryan languages were not, however, originally native to India; they were introduced by Aryan invaders from the north. Several scholars have demonstrated that pre-Indo-Aryan and pre-Dravidian bilingualism in India provided conditions for the far-reaching influence of Dravidian on the Indo-Aryan tongues in the spheres of phonology (*e.g.*, the retroflex consonants, made with the tongue curled upward toward the palate), syntax (*e.g.*, the frequent use of gerunds, which are nonfinite verb forms of nominal character, as in "by the falling of the rain"), and vocabulary (a number of Dravidian loanwords apparently appearing in the Rigveda itself).

Dravidian features in the Rigveda

Thus a form of Proto-Dravidian, or perhaps Proto-North Dravidian, must have been extensive in northern India before the advent of the Aryans. Apart from the survival of some islands of Dravidian speech, however, the process of replacement of the Dravidian languages by the Aryan tongues was entirely completed before the beginning of the Christian Era, after a period of bilingualism that must have lasted many centuries. Finally, the almost universal adoption of Indo-Aryan in the north and of Dravidian in the south has covered up the original linguistic diversity of India.

The circumstances of the advent of Dravidian speakers in India are shrouded in mystery. There are vague linguistic and cultural ties with the Urals, with the Mediterranean area, and with Iran. It is possible that a Dravidian-speaking people that can be described as dolichocephalic (longheaded from front to back) Mediterraneans mixed with brachycephalic (short-headed from front to back) Armenoids and established themselves in northwestern India during the 4th millennium BC. Along their route, these immigrants may have possibly come into an intimate, prolonged contact with the Ural-Altaic speakers, thus explaining the striking affinities between the Dravidian and Ural-Altaic language groups. Between 2000 and 1500 BC, there was a fairly constant movement of Dravidian speakers from the northwest to the southeast of India, and about 1500 BC three distinct dialect groups probably existed: Proto-North Dravidian, Proto-Central Dravidian, and Proto-South Dravidian. The beginnings of the splits in the parent speech, however, are obviously earlier. It is possible that Proto-Brāhūī was the first language to split off from Proto-Dravidian, probably during the immigration movement into India sometime in the 4th millennium BC, and that the next subgroup to split off was Proto-Kurukh-Malto, sometime in the 3rd millennium BC (see the family tree diagrams, Figures 20–22).

Possible Dravidian and Ural-Altaic contact

Compared to the work done on other language families, the progress in comparative Dravidian studies has been slow and firm results are still meagre. Considerable knowledge has been acquired in comparative phonology (sound systems), but correspondences have been worked out only for the sounds in the roots of words. Very little comparative work has been done on grammatical processes, and complete historical grammars of the literary languages are still lacking. Hence the reconstruction of any feature of the Dravidian protolanguage, with the possible exception of some parts of the phonology, must necessarily be considered very tentative.

The vowel system of Proto-Dravidian consisted of five vowels—**i, *u, *e, *o, *a* (an asterisk denotes an unattested, reconstructed, hypothetical form)—each having

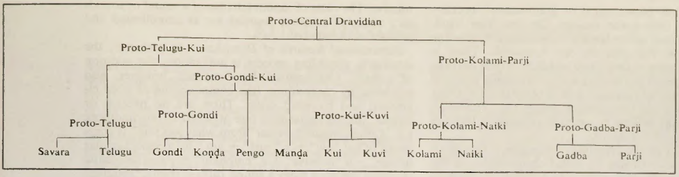


Figure 21: Central Dravidian subfamily.

two quantities, short and long. Relative stability of root vowels seems to have been the rule. The Proto-Dravidian consonant system consisted of obstruants (stops) *p, *t, *ṯ, *ṭ, *c, *k; nasals *m, *n, *ṇ, *ñ; laterals *l, *ḷ; the flap *r; the voiced retroflex continuant *ṛ; and the semivowels *y and *v. The most characteristic feature of the consonantal system was the six positions of articulation for obstruants: labial (with the lips), dental (tongue touching the back of the upper teeth), alveolar (tongue touching the upper gum ridge), retroflex (tip of tongue curled upward toward the palate and back), palatal (body of tongue touching the palate, or roof of the mouth), and velar (back of tongue touching the velum, or soft palate). The retroflex series was very distinctive and important and comprised an obstruant *ṭ, a nasal *ṇ, a lateral *ḷ, and a continuant *ṛ. No consonant of the alveolar or retroflex series began a word. In the final position all of the consonants occurred, but all of the obstruants were followed by an automatic release sound, the vowel *-u. Initial consonant clusters did not occur. There was only one series of obstruant phonemes (distinctive sounds); these sounds were voiceless (produced without vibration of the vocal cords) initially and voiced (with vocal cord vibration) between vowels. All Proto-Dravidian roots were monosyllables.

Proto-Dravidian used only suffixes, never prefixes or infixes, in the construction of inflected forms. Hence, the roots of words always occurred at the beginning. Nouns, verbs, and indeclinable words constituted the original word classes.

Process of Dravidian acculturation in India

During the 1st millennium BC, while Aryanization steadily progressed in north India, the Dravidian-speaking newcomers began to mix with the Negritos and Proto-Australoids in the south; this process of acculturation continued during the period from approximately 1200 to 600 BC. A movement of the Aryans into the south of India

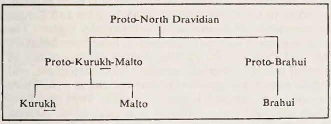


Figure 22: North Dravidian subfamily.

began sometime about 1000 BC. Before the 5th century BC, Proto-South Dravidian was probably still one language, but with two strongly marked dialects. Within Proto-Central Dravidian, a similarly deep two-way division also occurred, and as discussed above, North Dravidian must by that time have already been split into the Kurukh-Malto and Brahui subgroups (see the family tree diagrams, Figures 20–22).

Apart from a possible Dravidian word in the Hebrew text of the Bible (*tukkhiyīm* "peacocks"; *cf.* Tamil *tōkai* "tail of a peacock"), the Dravidian languages enter history in Sanskrit and Greco-Roman texts. The *Cēras,* a south Indian dynasty, are possibly mentioned in the early Sanskrit text *Aitareya Āraṇyaka.* Kātyāyana, a grammarian of the 4th century BC, mentions the countries of Pāṇḍya (Tamil *pāṇṭiya*), Cōla (Tamil *cōla*), and Kerala, or Cēra (Tamil *cēra*); these lands were well known to Kauṭilya (4th century BC), the author of the earliest treatise on statecraft, and mentions of them also appear in the edicts of the great Buddhist leader Aśoka (3rd century BC). The term *drāviḍa* itself is almost certainly a Sanskritization (with an inserted "hypercorrect" *r*) of the earlier Pāli and Prākrit terms *dāmiḷo, damiḷa, dāviḍa,* which must have been derived from the Tamil name of the language, *tamil.* A number of South Dravidian words, almost all of them geographic and dynastic names, occur in such Greco-Roman sources as the *Periplus maris Erythraei* ("Circumnavigation of the Erythraean Sea") of about AD 89 and in the writing of Ptolemaeus of Naukratis of the 2nd century AD; it is also very probable that Western-language terms for rice (compare Italian *riso,* Latin *oryza,* Greek *oryza*) and ginger (compare Italian *zenzero,* German *Ingwer,* Greek *zingiberis*) are cultural loans from Old Tamil, in which they are *arici* and *iñcivēr,* respectively.

Sometime during the reign of Aśoka (3rd century BC), the two South Dravidian languages, Tamil and Kannada, developed into distinct idioms and the two cultures emerged as separate entities; a third major Dravidian linguistic and cultural unit, Telugu, appeared in the Andhra country. In the period from 300 to 100 BC, one of the pre-Tamil dialects (probably that of Madurai) gained prestige and became the standard literary language (*centamil*), the written form of early Old Tamil, which became established in poetic texts and in its earliest grammar, *Tolkāppiyam.* During the same period, about 250 BC, the Aśokan Southern Brāhmī script was adapted for Tamil and was used in short cave inscriptions by Jain monks over a period of several centuries, dating approximately from the 2nd century BC to the 5th century AD.

Early inscriptions and writings

The earliest inscriptions in Kannada may be dated at AD 450; Kannada literature begins with Nṛpatuṅga's *Kavirājamārga,* about AD 850. The oldest Telugu inscription is from AD 633, and the literature begins with the grammarian Nannaya's 11th-century translation of the Sanskrit classic the *Mahābhārata.* In Malayalam, the earliest writings are from the close of the 9th century, and the first literary text is probably the *Bhāṣākauṭalīyam,* AD 1125–1250.

Since these attested beginnings, the four languages—Tamil, Malayalam, Kannada, and Telugu—have been used continuously in administration and literature up to the present day. In addition to possessing an immense wealth of epigraphic and literary texts, they all developed pronounced features of diglossia, a dichotomy between the standardized, formal language and the informal, colloquial speech, which is divided into regional as well as social dialects. In modern times, all of the four cultivated languages have adapted quickly to new conditions resulting from economic, social, and political changes. All of these languages are used in teaching basic courses in science and the arts; and new technological terminology is coined, sometimes based either on English or Sanskrit models, but often on exclusively indigenous linguistic material (in Tamil).

To date, nothing is known about the history of the nonliterary Dravidian languages before their "discovery," which began at the end of the 18th century. The Gonds, however, are mentioned (as Gondaloi) by Ptolemy of Naukratis, writing in the 2nd century AD.

A tendency toward structural and systemic balance and stability is characteristic of the Dravidian group. Nevertheless, there is no doubt about the influence of the other languages of India. Dravidian languages show extensive lexical (vocabulary) borrowing, but only a few traits of structural (either phonological or grammatical) borrowing, from the Indo-Aryan tongues. On the other hand, Indo-Aryan shows rather large-scale structural borrowing from Dravidian, but relatively few loanwords. There is indeed a possibility of Dravidian and Indo-Aryan drawing even closer together in the future; but it is highly doubtful that a new family of languages will develop in such a way that the bases of the contributing groups (*i.e.*, Dravidian and Indo-Aryan) will be completely eliminated through the phenomena of borrowing.

Agglutinating character of Dravidian

Characteristics of the Dravidian languages. Dravidian languages would probably be called agglutinative in the categorization of the 19th-century philologists. An agglutinative language incorporates separate formal units of distinct meaning into a single word. There are some elements of "internal flexion" (*e.g.*, the alternation of short–long root vowels in derived words), however, as well as regular alternations in vowel and consonant quantities within the root. Relatively low receptivity to change results in a slower rate of change than is found in the Indo-European language family.

The degree of phonetic divergence among the Dravidian languages is not very great; hence, etymologies are not too difficult to discover. The territory occupied by Dravidian speakers in India may be characterized as a large dialect area resembling the area of the Romance languages, with numerous boundaries marked by bundles of isoglosses (an isogloss is a boundary line that separates the areas of two differing features of language usage), but also with many isoglosses enclosing more than one language. In any study of Dravidian, therefore, both evolution and diffusion must be taken into account.

Sounds of Dravidian. Compared to the reconstructed system of Proto-Dravidian phonemes (distinctive sounds), the most striking developments in vowels are the gradual elimination of the contrast between *e* and *ē* (long *e*) and *o* and *ō* (long *o*) in Brahui, as a result of the influence of Indo- Aryan languages or Iranian or both; the raising of Proto-Dravidian **e* and **o* to *i* and *u* and the lowering of these protolanguage sounds in Brahui; and the merger of Proto-Dravidian **i* and **u* with **e* and **o* in the South Dravidian languages before a consonant plus the vowel *a*. Also noteworthy are the emergence of retroflex vowels (*i.e.*, centralized vowels "coloured" by neighbouring retroflex consonants) in Kodagu and Irula; the nasalization of vowels, as in colloquial Tamil; the loss of vowels in unaccented noninitial syllables in Toda, Kota, some dialects of Kannada, and Tamil, and the resulting consonant clusters (*e.g.*, Kota *anjrčgčgvdk*, "because of the fact that [someone] will cause [someone] to terrify [someone]"). Metathesis (the transposition of sounds, as in "aks" from "ask") and vowel contraction resulted in initial consonant clusters in Telugu and other Central Dravidian languages—*e.g.*, Tamil *koḻu*, but Kui *krōga*, both meaning "fat."

Consonant changes

Among the most important consonantal developments are the loss of **c*-, a typical South Dravidian phenomenon that seems to be still in progress (*e.g.*, Proto-Dravidian **cā̆ṛ*-, but Tamil *alal* "to burn," and *talal* "to glow"); the velarization of **c*- to *k*- in North Dravidian when the sound is followed by ū̆ (*e.g.*, Tamil *cuṭu* "be hot," but Malto *kut*- "burn"); the palatalization of Proto-Dravidian **k*- to *c*- before front vowels in Tamil, Malayalam, and Telugu (*e.g.*, **ke*- "red," but Tamil *ce*-); and the replacement of **k*- in North Dravidian by *x* before ā̆, ō̆, and ū̆ (*e.g.*, Tamil *kal*, but Brahui *xal*, "stone"). The retroflex voiced continuant **ṛ* has been preserved only in the old stages of the cultivated languages and partly in modern Tamil and Malayalam; elsewhere, it merged with *ḷ*, *ḍ*, and other sounds. Some languages, notably Kannada, developed a secondary *h*-, not inherited from the parent speech (*e.g.*, Tamil *peyar*, Old Kannada *pesar*, but Modern Kannada *hersru*, "name"). According to the Dravidian scholar Bhadriraju Krishnamurti, a laryngeal (or *h*-type of sound) should be reconstructed for some items in Proto-Dravidian.

Problems of accent and intonation still remain to be worked out. Word stress is predictable, always occurring on the radical (initial) syllable and therefore being nondistinctive. The rules of sandhi (change of a sound or sounds as a result of adjacent sounds) are as complicated and delicate as in Sanskrit.

Grammatical features of Dravidian. In grammar, the absolutely prevailing process is suffixation, the addition of suffixes. Grammatical functions are, however, also expressed by composition (the compounding of word elements) and by word order. There are no prefixes or infixes. Suffixes agglutinate (are attached to one another); *e.g.*, Tamil *coṉṉatilēyiruntu* "from what was said" is composed of *col* "say" + *n* "past" + *atu* "3rd person singular neuter" + *il* "locative" + *ē* "emphatic" + *y* (an automatic insertion resulting from a sound rule) + *iruntu* "ablative" (*iruntu* comes from *iru* "be" + *nt/u* "past").

The major word classes are nouns (substantives, numerals, pronouns), adjectives, verbs, and indeclinables (particles, enclitics, adverbs, interjections, onomatopoetic words, echo words). There are two numbers and four different gender systems, the "original" probably having "male: non-male" in the singular and "person:non-person" in the plural. The pronoun has a category "inclusive:exclusive" in the 1st person plural. A characteristic derivation is that of "pronominalized" or "personal" nouns and adjectives; *e.g.*, Old Tamil *iḷai* "youth," *iḷai-y-am* "young-we," *iḷai-y-ar* "young-they."

Finite forms of the verb (forms showing person and number) are, ultimately, "pronominalized" verb stems; *e.g.*, Tamil *aṭi-(y)-ēn* ("slave"—1st person singular) "I am a slave"; *nal-(l)-ēn* ("good"—1st person singular) "I am good"; *pō-v-ēn* ("go"—future—1st person singular) "I shall go." The most characteristic feature of the Dravidian verb is a full-fledged negative system: all of the positive verb forms have their corresponding negative counterparts. Verbs are intransitive, transitive, and causative; there are also active and passive forms. The main (and probably original) dichotomy in tense is past:non-past. Present tense developed later and independently in each language or subgroup.

Negative verbs

In a sentence, however complex, only one finite verb occurs, normally at the end, preceded if necessary by a number of gerunds. Gerunds and participles, as well as verb-nouns, play an important role. The determining member always precedes the determined; *e.g.*, Tamil *poṉ* "gold" + *nakaram* "city" becomes *poṉṉakaram* "city of gold, golden city." Word order follows certain basic rules but is relatively free.

Vocabulary. In vocabulary, different Dravidian languages were receptive to loanwords in differing degrees. Among the cultivated languages, Tamil has the relatively lowest number of Indo-Aryan loanwords (18–25 percent, according to the style), whereas in Malayalam and Telugu the percentage of loanwords is substantially higher. The most important sources of loanwords have been Sanskrit, Pāli, and Prākrit (with varying degrees of importance in different periods); in modern times Urdu, Portuguese, and English have made significant contributions as well. There was only very limited lexical borrowing from one Dravidian language into another in historical times. Among all of the Dravidian languages, Brahui, in Pakistan, is inevitably the one most influenced by Indo-Aryan and Iranian; in contrast, Toda is probably the one language least influenced by any other idiom. In Tamil, there is currently a very notable and active purifying movement; it aims at removing as many borrowed "Sanskritic" (but not English) vocabulary items as possible. Such purism has not yet occurred in any other of the cultivated Dravidian languages.

Writing. Writing was first developed in Tamil Nadu, sometime about 250 BC, when the Aśokan Southern Brāhmī script was adapted for Tamil. The earliest inscriptions in Tamil script proper are the Pallava copperplates of about AD 550. The Kannada–Telugu script is based on Cālukya (6th century) inscriptions; the Grantha script, used in Tamil Nadu for Sanskrit since the 6th century, was accommodated for Malayalam and Tulu. Apart from these, Tamil has an old cursive script called Vaṭṭeḻuttu, "round script," and Malayalam possesses its own modern cursive form, Koleḻuttu, "rod-script."

(K.V.Z.)

AUSTROASIATIC LANGUAGES

Austroasiatic languages are spoken by more than 65 million people scattered throughout Southeast Asia and eastern India. The stock comprises about 150 languages, most of them having numerous dialects. Khmer, Mon, and Vietnamese are culturally the most important and have the longest recorded history. The rest are languages of nonurban minority groups written, if at all, only recently. The stock is of great importance as a linguistic substratum for all Southeast Asian languages.

Superficially, there seems to be little in common between a monosyllabic tone language such as Vietnamese and a polysyllabic toneless Muṇḍā language such as Muṇḍārī of India; linguistic comparisons, however, confirm the underlying unity of the family. The date of separation of the two main Austroasiatic subfamilies—Muṇḍā and Mon-Khmer—has never been estimated and must be placed well back in prehistory. Within the Mon-Khmer subfamily itself, 12 main branches are distinguished; glottochronological estimates of the time during which specific languages have evolved separately from a common source indicate that these 12 branches all separated about 3,000 to 4,000 years ago.

Main subfamilies

Relationships with other language families have been proposed, but, because of the long durations involved and the scarcity of reliable data, it is very difficult to present a solid demonstration of their validity. In 1906 Wilhelm Schmidt, a German anthropologist, classified Austroasiatic together with the Austronesian family (formerly called

Table 41: Austroasiatic languages

Austroasiatic stock	areas where spoken*
MON-KHMER FAMILY	
Khasian branch	Meghālaya (NE India)
Khāsī, Synteng, Lyng-ngam	
Amwi (War)	
Palaungic branch (Palaung-Wa)	
Kano' (Danau)	NE Myanmar
Palaung-Riang subbranch	NE Myanmar, SW China
Ta-ang (Palaung, Gold Palaung), Ka-ang	
Da-ang (Pale, Silver Palaung)	
Na-ang, Ra-ang	
Riang, White-striped Riang, Black Riang	
Angkuic subbranch	
Angku (Kon-Keu), U, Hu	SW China, NE Myanmar
Mok, Man-Met	NE Myanmar, SW China, N Thailand
Khabit	NW Laos
Samtao of Laos	NW Laos
Lamet (Khamet), Ramet (Lua')	NW Laos, N Thailand
Waic subbranch	
Plang (Bu Lang, Samtao of Myanmar)	SW China, NE Myanmar
Wa, Paraok, Avüa, Alva	SW China, NE Myanmar
Phalok	N Thailand
Lawa (Ravüa, Lua')	N Thailand
Mang	N Vietnam
Khmuic branch	
Khmu (Kammu, Xa Khmu), Yuan	N Laos, N Thailand
Mal (Thin, Prai, Phai, Lua')	NW Laos, N Thailand
Mlabri, Yumbri	N Thailand
Iduh (Odu, Thai Hat)	NE Laos, NW Vietnam
Thai Then	N Laos
Phong, Kaniang, Piat, Phong Lan	NE Laos
Khsing Mul (Puoc, Ksing Mun)	NE Laos, NW Vietnam
Khang	NW Vietnam
Pakanic branch	S China
Palyu (Bolyu, Lai)	
Pakan (Bugan)	
Vietic branch	
Viet-Muong subbranch	
Vietnamese (Kinh)	Vietnam, S China
Muong, Nguon	N Vietnam
Arem	NW Vietnam
Sach, May, Ruc	NW Vietnam
Thavung, Ahlau, Aheu (Phone Soung)	C Laos
Maleng (Pakatan), Malieng	C Laos, NW Vietnam
Tum, Cuoi, Pong, Uy-Lo, Khong-Kheng	NW Vietnam, C Laos
Katuic branch	
West Katuic subbranch	
Bru, Makong, Kanay	C Vietnam, C Laos, NE Thailand
So, Tri (Chali), Truy	C Laos, NE Thailand
Kuay (Souei, Kuy), Yeu	NE Thailand, S Laos, N Cambodia
East Katuic subbranch	
Katu, Kantu, Phuong	C Vietnam, C Laos
Pacoh	C Vietnam, C Laos
Ngkriang (Ngeq)	C Laos
Katang	C Laos
Ta-oih (Ta-oi, Ta-uas), Ong, Yir	C Laos
Bahnaric branch	
West Bahnaric subbranch	
Brao (Lave), Krung, Kravet	S Laos, NE Cambodia
Jru' (Loven)	S Laos
Nyah Heuny (Ngaheune)	S Laos
Sok, Oy, Sou, Cheng, Sapuan	S Laos
Northwest Bahnaric subbranch	
Tarieng (Talieng)	S Laos
Alak (Harlaak), Lawi	S Laos
North Bahnaric subbranch	
Kacho'	NE Cambodia
Rengao	C Vietnam
Sedang (Hatea), Tadrah, Didrah	C Vietnam
Hre	C Vietnam
Jeh, Halang, Kayong	C Vietnam
Cua, Takua, Duan	C Vietnam
Central Bahnaric subbranch	
Bahnar	C Vietnam
Tampuan	NE Cambodia
South Bahnaric subbranch	
Mnong, Biat, Phnong	S Vietnam, SE Cambodia
Sre (Koho), Maa'	S Vietnam
Stieng	SE Cambodia
Chrau	S Vietnam
Pearic branch	
Chong	SE Thailand
Chung (Sa-och)	W Cambodia
Song of Trat	SE Thailand
Samre (Eastern Pear)	SE Thailand, W Cambodia
Samrai (Western Pear)	W Cambodia
Song of Kampong Spoe	C Cambodia
Pear of Kampong Thum	N Cambodia
Khmeric branch	Cambodia, NE and SE Thailand, S Vietnam
Khmer, Northern Khmer, Southern Khmer, Western Khmer	
Old Khmer (Angkorian), Pre-Angkorian Old Khmer	
Monic branch	
Mon	C and S Myanmar; N, W, and C Thailand
Old Mon	C Myanmar; C, N, and NE Thailand
Nyah Kur (Chao Bon)	C and NE Thailand
Aslian branch	
North Aslian subbranch (Semang)	
Kenta', Kensiw, Ten-en	S Thailand, NW Malaysia
Jahai	N Malaysia
Menriq	N Malaysia
Bateg	N and C Malaysia
Che' Wong (Siwang)	C Malaysia
Senoic subbranch (Sakai)	
Lanoh, Semnam, Sabum	NW Malaysia
Temiar	C Malaysia
Semai	C Malaysia
Jah Hut (Jah Het)	C Malaysia
South Aslian subbranch (Semelaic)	
Betise' (Mah Meri, Besisi)	S Malaysia
Semelai	S Malaysia
Semaq Beri	S Malaysia
Nicobarese branch	Nicobar Islands (India)
Car, Chowra, Teresa, Bompaka	
Nancowry (Central Nicobar), Camorta, Trinkat, Katchall	
Coastal Great Nicobar, Little Nicobar	
Shompe	
MUṆḌĀ FAMILY	E India
North Muṇḍā subfamily	
Korkū	Madhya Pradesh
Kherwārī branch	Bihār, Bengal, Orissa
Santhālī	
Muṇḍārī	
Ho, Bhumij	
South Muṇḍā subfamily	
Central Muṇḍā branch	Orissa, Bihār
Khaṛiā	
Juāng	
Korāput Muṇḍā branch	Orissa, Andhra Pradesh
Gutob, Remo	
Sora (Savara), Juray, Gorum	

*Capital letters denote direction, C standing for central.

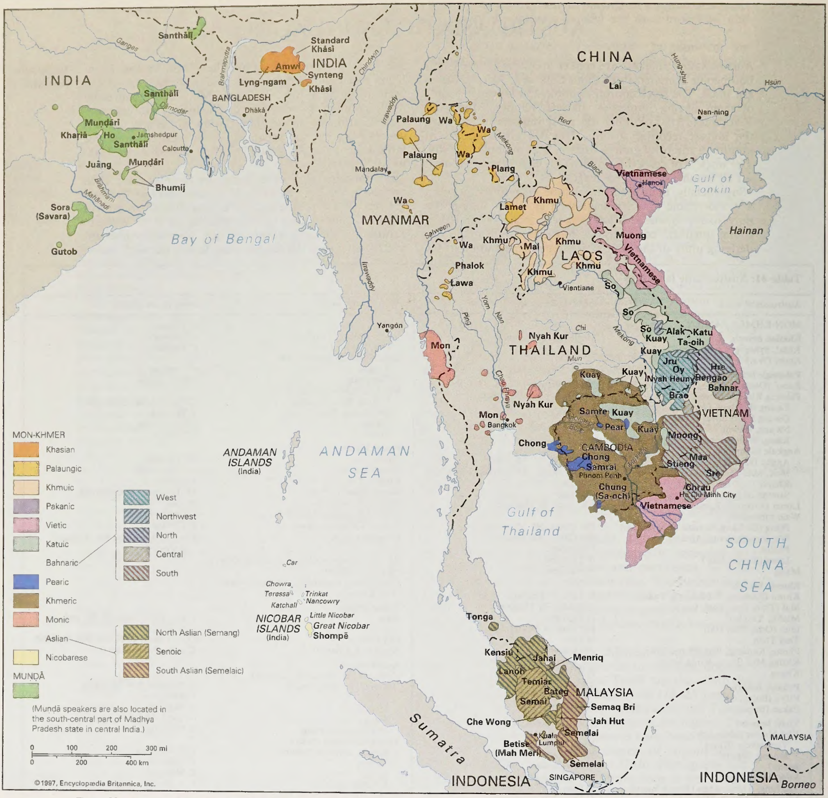


Figure 23: Distribution of the Austroasiatic languages.

Malayo-Polynesian) to form a larger family called Austric. Paul K. Benedict, an American scholar, extended the Austric theory to include the Tai-Kadai family of Southeast Asia and the Miao-Yao (Hmong-Mien) family of China, together forming an "Austro-Tai" superfamily.

Regarding subclassification within Austroasiatic, there have been several controversies. Schmidt, who first attempted a systematic comparison, included in Austroasiatic a "mixed group" of languages containing "Malay" borrowings and did not consider Vietnamese to be a member of the family. On the other hand, some of his critics contested the membership of the Muṇḍā group of eastern India. The "mixed group," called Chamic, is now considered to be Austronesian. It includes Cham, Jarai, Rade (Rhade), Chru, Roglai, and Haroi and represents an ancient migration of Indonesian peoples into southern Indochina. As for Muṇḍā and Vietnamese, the works of the German linguist Heinz-Jürgen Pinnow on Khariā and of the French linguist André Haudricourt on Vietnamese tones have shown that both language groups are Austroasiatic.

CLASSIFICATION OF THE AUSTROASIATIC LANGUAGES

The work of classifying and comparing the Austroasiatic languages is still in the initial stages. In the past, classification was done mainly according to geographic location. For instance, Khmer, Pear, and Stieng, all spoken on Cambodian territory, were all lumped together, although they actually belong to three different branches of the Mon-Khmer subfamily.

Khmer and Vietnamese are the most important of the Austroasiatic languages in terms of numbers of speakers. They are also the only national languages—Khmer of Cambodia, Vietnamese of Vietnam—of the Austroasiatic stock. Each is regularly taught in schools and is used in mass media and on official occasions. Speakers of most other Austroasiatic languages are under strong social and political pressure to become bilingual in the official languages of the nation in which they live. Most groups are too small or too scattered to win recognition, and for many the only chance of cultural survival lies in retreating to a mountain or jungle fastness, a strategy that reflects long-standing Austroasiatic tradition.

Khmer and Vietnamese: national languages

LINGUISTIC CHARACTERISTICS

Phonological characteristics. The sound systems of Austroasiatic languages are fairly similar to each other, but Vietnamese and the Muṇḍā languages, under the influence of Chinese and Indian languages respectively, have diverged considerably from the original type. The usual Austroasiatic word structure consists of a major syllable sometimes preceded by one or more minor syllables. A minor syllable has one consonant, one minor vowel, and optionally one final consonant. Most languages have only one possible minor vowel, but some have a choice of three (*e.g.*, *a, i,* or *u*) or even use vocalic nasals (*m* or *n*) and liquids (*l* or *r*) as minor vowels. Major syllables are composed of one or two initial consonants, followed by one major vowel and one final consonant. Many languages—*e.g.*, Khmer, Mon, and Bahnar—allow major syllables without final consonants, but no Austroasiatic language allows combinations of two or more final consonants.

Consonants. A typical feature of Mon-Khmer languages, uncommon in the Muṇḍā subfamily, is to allow a great variety of two-consonant combinations at the beginning of major syllables. Khmer is especially notable for this. At the end of a word, the inventory of possible consonants is always smaller than at the beginning of the major syllable and is considerably smaller when contact with Tai-Kadai or Sino-Tibetan languages has been extensive. These two properties combine to give Mon-Khmer words their characteristic rhythmic pattern, rich and complicated at the beginning, simple at the end.

Several Mon-Khmer languages—*e.g.*, Khmer, Katu, Mon, and some forms of Vietnamese—allow implosive ɓ and ɗ at the beginning of major syllables. These sounds, pronounced with a brief suction of the air inward, have sometimes been called pre-glottalized, or semi-voiceless, sounds. They probably existed in the ancestral language called Proto-Mon-Khmer but have disappeared in many modern languages.

A series of aspirated consonants, p^h, t^h, c^h, and k^h, pronounced with a small puff of air, is found in several branches or subbranches of Mon-Khmer (Pearic, Khmuic, South Aslian, Angkuic), but this is not a typical feature of the family, and it probably did not exist in the ancestral language.

Most Austroasiatic languages have palatal consonants (*č* or *ñ*) at the end of words; they are produced with the blade of the tongue touching the front part of the palate. Austroasiatic languages stand apart from most other languages of Asia in having final consonants of this type.

Great variety of vowels

Vowels. Typical of Mon-Khmer languages is an extraordinary variety of major vowels: systems of 20 to 25 different vowels are quite normal, while several languages have 30 and more. Nasal vowels are sometimes found, but in any one language they do not occur very frequently. Four degrees of height are usually distinguished in front and back vowels, as well as in the central area. The variety of Khmer spoken in Surin (Thailand) distinguishes five degrees of height, plus diphthongs, all of which can be either short or long, for a total of 36 major vowels.

Tones. Most Austroasiatic languages, notably Khmer, Mon, Bahnar, Kuay, and Palaung, do not have tones. This is noteworthy, considering that the language families found to the north—Tai-Kadai, Sino-Tibetan, and Hmong-Mien (Miao-Yao)—all have tones. The few Austroasiatic languages that are tonal—*e.g.*, Vietnamese, the Angkuic subbranch, and the Pakanic branch—are found in the northern geographic range of the family. They have acquired tones independently from each other, in the course of their own history, as a result of contact and bilingualism with language families to the north. Tones are not posited for any ancient stage of Mon-Khmer or Austroasiatic.

Registers. Much more characteristic of the Austroasiatic stock is a contrast between two or more series of vowels pronounced with different voice qualities called registers. The vowels may have, for example, a "breathy" register, a "creaky" register, or a clear one. This feature, which is fairly rare the world over, is found, for example, in Mon, Wa, and Kuay, which distinguish breathy from clear vowels; in some Katuic languages, which distinguish creaky vowels from clear ones; and in the Pearic branch, which cumulates both distinctions. These registers have a variety of historical origins; for some languages (such as Mon) they are a fairly recent innovation, but for others (such as Pearic) they may be very ancient, perhaps dating to the ancestral language called Proto-Austroasiatic.

Grammatical characteristics. *Morphology.* In morphology (word formation), Muṇḍā and Vietnamese again show the greatest deviations from the norm. Muṇḍā languages have an extremely complex system of prefixes, infixes (elements inserted within the body of a word), and suffixes. Verbs, for instance, are inflected for person, number, tense, negation, mood (intensive, durative, repetitive), definiteness, location, and agreement with the object. Furthermore, derivational processes indicate intransitive, causative, reciprocal, and reflexive forms. On the other hand, Vietnamese has practically no morphology.

Common features

Between these two extremes, the other Austroasiatic languages have many common features. (1) Except in Nicobarese, there are no suffixes. A few languages have enclitics, certain elements attached to the end of noun phrases (possessives in Semai, demonstratives in Mnong), but these do not constitute word suffixes. (2) Infixes and prefixes are common, so that only the final vowel and consonant of a word root remain untouched. It is rare to find more than one or two affixes (*i.e.*, prefixes or infixes) attached to one root; thus, the number of syllables per word remains very small. (3) The same prefix (or infix) may have a wide number of functions, depending on the noun or verb class to which it is added. For instance, the same nasal infix may turn verbs into nouns and mass nouns into count nouns (noun classifiers). (4) Many affixes are found only in a few fossilized forms and often have lost their meaning. (5) Expressive language and wordplay are embodied in a special word class called "expressives." This is a basic class of words distinct from verbs, adjectives, and adverbs in that they cannot be subjected to logical negation. They describe noises, colours, light patterns, shapes, movements, sensations, emotions, and aesthetic feelings. Synesthesia is often observable in these words and serves as a guide for individual coinage of new words. The forms of the expressives are thus quite unstable, and the additional effect of wordplay can create subtle and endless structural variations.

Syntax. In syntax, possessive and demonstrative forms and relative clauses follow the head noun; if particles are found, they will be prepositions, not postpositions (elements placed after the word to which they are primarily related), and the normal word order is subject–verb–object. There is usually no copula equivalent to the English verb "be." Thus, an equational sentence will consist of two nouns or noun phrases, separated by a pause. Predicates corresponding to the English "be + adjective" usually consist of a single intransitive (stative) verb. Ergative constructions (in which the agent of the action is expressed not as the subject but as the instrumental complement of the verb) are quite common. Also noteworthy are sentence final particles that indicate the opinion, the expectations, the degree of respect or familiarity, and the intentions of the speaker. Muṇḍā syntax, once again, is radically different, having a basic subject–object–verb word order, like the Dravidian languages of India. It is quite conceivable that the complexity of Muṇḍā verb morphology is a result of the historical change from an older subject–verb–object to the present subject–object–verb basic structure.

Vocabulary. The composition of the vocabulary of the Austroasiatic languages reflects their history. Vietnamese, Mon, and Khmer, the best-known languages of the family, came within the orbit of larger civilizations and borrowed without restraint—Vietnamese from Chinese, Mon and Khmer from Sanskrit and Pāli. At the same time, they have lost a large amount of their original Austroasiatic vocabulary. It is among isolated mountain and jungle groups that this vocabulary is best preserved. But other disruptive forces are at work there. For instance, animal names are subject to numerous taboos, and the normal name is avoided in certain circumstances (*e.g.*, hunting, cooking, eating, and so on). A nickname is then invented, often by using a kinship term ("Uncle," "Grandfather")

Word taboos

followed by a pun or an expressive adverb describing the animal. In the course of time, the kinship term is abbreviated (thus many animal names begin with the same letter), the normal name is forgotten, and the nickname becomes standard. As such, it is then in turn avoided, and the process is repeated. There are also taboos on proper names; *e.g.,* after a person's death, his name and all words that resemble it are avoided and replaced by metaphors or circumlocutions. These replacements may explain why, for instance, the Nicobarese languages, which seem closely related, have few vocabulary items in common. In general, new words and fine shades of meanings can always be introduced by wordplay and from the open-ended set of expressive forms. Borrowings from the nearest majority languages are also common.

Writing systems and texts. Two Austroasiatic languages have developed their own orthographic systems and use them to this day. For both scripts, the letter shapes and principles of writing were borrowed from Indian alphabets (perhaps those of the Pallava dynasty in South India) that were in use in Southeast Asia at the time. Both Austroasiatic groups modified these alphabets in their own way, to suit the complex phonology of their languages. The most ancient inscriptions extant are in Old Mon and Old Khmer in the early 7th century. The monuments of Myanmar (Burma), Thailand, and Cambodia have preserved a large number of official inscriptions in these two languages. Both alphabets were in turn used as models by other peoples for writing their own languages, the Thai speakers using Khmer letters and the Burmese speakers using Mon letters. The religious literature in Old and Middle Mon played a very important role in the spreading of Theravāda Buddhism to the rest of Southeast Asia.

Because Vietnam was a Chinese province for a thousand years, the Chinese language was used and written there for official purposes. In the course of time (perhaps as early as the 8th century AD), a system called Chunom (popular writing) was developed for writing Vietnamese with partly modified Chinese characters. About 1650, Portuguese missionaries devised a systematic spelling for Vietnamese, based on its distinctive sounds (phonemes). It uses the Latin (Roman) alphabet with some additional signs and several accents to mark tones. At first, and for a long time, the use of this script was limited to Christian contexts, but it spread gradually, and in 1910 the French colonial administration made its use official. Now called quoc-ngu (national language), it is learned and used by all Vietnamese.

Most other Austroasiatic languages have been written for less than a century; the literacy rate remains very low with a few exceptions (*e.g.,* Khāsī). Dictionaries and grammars have been written only for the most prominent languages, with traditional and often insufficient methods. Many languages have only been described briefly in a few articles, and many more are little more than names on the map. (G.Di.)

SINO-TIBETAN LANGUAGES

In the narrowest sense, the Sino-Tibetan languages include the Chinese and Tibeto-Burman languages. In terms of numbers of speakers, they comprise the world's second largest language family (after Indo-European), including more than 300 languages and major dialects. In a wider sense, Sino-Tibetan has been defined as also including the Tai (Daic), Karen, and Miao-Yao (Hmong-Mien) languages and even the Ket language in central Siberia (the latter affiliation seems untenable). Some linguists connect the Mon-Khmer family of the Austroasiatic stock or the Austronesian (Malayo-Polynesian) family, or both, with Sino-Tibetan; a suggested term for this most inclusive group, which seems to be based on premature speculations, is Sino-Austric. Other scholars see a relationship of Sino-Tibetan with the Athabascan and other languages of North America, but proof of this is beyond reach at the present state of knowledge.

Figure 24: Distribution of the Sino-Tibetan languages.

Sino-Tibetan languages were known for a long time by the name of Indochinese, which is now restricted to the languages of Vietnam, Laos, and Cambodia. They were also called Tibeto-Chinese until the now universally accepted designation Sino-Tibetan was adopted. The term Sinitic also has been used in the same sense, but also as below for the Chinese subfamily exclusively. (In the following discussion of language groups, the ending *-ic,* as in Sinitic, indicates a relatively large group of languages, and *-ish* denotes a smaller grouping.)

Distribution and classification of Sino-Tibetan languages

DISTRIBUTION

Areas in which the languages are spoken

Sinitic languages. Sinitic languages, commonly known as the Chinese dialects, are spoken in China and on Taiwan and by important minorities in all the countries of Southeast Asia (by a majority only in Singapore). In addition, Sinitic languages are spoken by Chinese immigrants in many parts of the world, notably in Oceania and in North and South America; altogether there are at least 1.14 billion speakers of the Chinese dialects. Sinitic is divided into a number of language groups, by far the most important of which is Mandarin (or Northern Chinese). Mandarin, which includes Modern Standard Chinese (based on the Peking dialect), is not only the most important language of the Sino-Tibetan family but also has the most ancient writing tradition still in use of any modern language. The remaining Sinitic language groups are Wu (including Shanghai dialect), Hsiang (Hunanese), Kan, Hakka, Yüeh (Cantonese, including Canton and Hong Kong dialects), and Min (including Fuchow, Amoy, Swatow, and Taiwanese).

Tibeto-Burman languages. Tibeto-Burman languages are spoken in Tibet and Myanmar (Burma); in the Himalayas, including the countries of Nepal and Bhutan and the state of Sikkim, India; in Assam, India, and in Pakistan and Bangladesh; they also are spoken by hill tribes throughout mainland Southeast Asia and central China (the provinces of Kansu, Tsinghai, Szechwan, and Yunnan). Tibetic (*i.e.,* Tibetan in the widest sense of the word) comprises a number of dialects and languages spoken in Tibet and the Himalayas. Burmic (Burmese in its widest application) includes Yi (Lolo), Hani, Lahu, Lisu, Kachin (Ching-p'o [Jingpo]), Kuki-Chin, the obsolete Hsi-hsia (Tangut), and other languages. The Tibetan writing system (which dates from the 7th century) and the Burmese (dating from the 11th century) are derived from the Indo-Aryan (Indic) tradition; the Hsi-hsia system (developed in the 11th–13th century in northwestern China) was based on the Chinese model. Pictographic writing systems, which show some influence from Chinese, were developed within the past 500 years by Yi and Na-hsi (Naxi, formerly Moso) tribes in western China. In modern times many Tibeto-Burman languages have acquired writing systems in Roman (Latin) script or in the script of the host country (Thai, Burmese, Indic, and others).

CLASSIFICATION

The old literary languages, Chinese, Tibetan, and Burmese, are generally considered as representatives of three major divisions within Sino-Tibetan (Sinitic, Tibetic, and Burmic, respectively). A fourth literary language, Thai, or Siamese (written from the 13th century), represents what was accepted for a long time as a Tai division of Sino-Tibetan or as a division of a Sino-Tai family (see below *Tai languages*). This relationship is now more commonly considered nongenetic in that most of the shared vocabulary is more likely attributable to a history of cultural borrowing than to derivation from a common ancestral language.

Sinitic stands apart from Tibetic and Burmic on many grounds, including vocabulary, morphology, syntax, and phonology. Most scholars agree on combining Tibetic and Burmic into a Tibeto-Burman subfamily, which also includes Bodo-Garo or Baric but not Karenic. If Karenic is to be considered Sino-Tibetan, it must be set up as an independent member of a Tibeto-Karen group that includes Tibeto-Burman. The special affinities between Sinitic and Karenic (especially in syntax) are then considered secondary. The two closely related language groups, Miao and Yao (also known as Hmong and Mien), are thought by some to be very remotely related to Sino-Tibetan; they are spoken in western China and northern mainland Southeast Asia and may well be of Austro-Tai stock.

Problems in classifying Karenic

In attempting to determine the exact interrelationship of the Tai languages, Karenic, Sino-Tibetan, and several marginal tongues, scholars must keep in mind that a discernible layer of Sino-Tibetan features in a given language may have been superimposed upon an older, non-Sino-Tibetan foundation (called the substratum language). Attributing a language to Sino-Tibetan or to another family may depend entirely on the ability of scholars to identify the substratum. Thus, if Tai is not considered as a division of Sino-Tibetan, it is because the substratum has been recognized as Austronesian; if Karen is still included among Sino-Tibetan languages on some level, it is perhaps because identification of a substratum is still lacking. Among the languages that have been classified as Sino-Tibetan, a great many are known only from word lists or have not yet been described in a way that makes valid comparisons possible.

A number of Sino-Tibetan languages are enumerated below together with their most likely affiliation. Some scholars believe the Tibetic and Burmic divisions to be premature and that for the present their subdivisions

Table 42: Tibetic (Bodic) Languages*

	areas where spoken	number of speakers†
Bodish-Himalayish		
Bodish languages		
Tibetan (with branches and dialects)	Tibet, Nepal, Bhutan, Pakistan, sporadically in India, the Chinese provinces of Kansu, Tsinghai, Szechwan, and Yunnan	4,890,000
Central group: Lhasa, Khams, Kagate, Jad, Nyamkat (Mnyamskad)	Nepal, China	
Southern group: Spiti, Sharpa, Sikkim, Lhoke	Bhutan, India, Nepal	
Northern group: Ambo (Ngambo), Chone	China	
Western group: Balti, Purik (Burig), Ladākhi (Ladwags)	India, Pakistan	
Derge	China	
Gurung	central Nepal	230,000
Gyarung (Rgyarung)	Tibet, Szechwan	100,000
Himalayish languages		118,000
Kanauri branch: Thebor, Bunan, Kanashi, Chitkhuli, Manchati, Rangloi, Chamba Lahuli	India	
Almora branch: Rangkas and others	Nepal	
Kirantish (Bahing-Vayu)		440,000
Eastern (Bahing) branch: Bahing, Sunwar, Dumi, Khambu, Rodong, Waiing, Lambichong, Lohorung, Limbu, Yakha	eastern Nepal	
Western (Vayu) branch: Vayu, Chepang; Magari (perhaps)	central Nepal	
Newari	central Nepal	550,000
Mirish (Mishingish)	Assam (India)	
Miri (Mishing)		
Abor		
Dafla (Nyising)		
Digaro (Taying)	Assam, Tibet	
Miju	Assam, Tibet	
Other Tibetic languages		
Hruso (Hurso, Aka)	northern Assam	
Dhimal	Darjeeling area (India)	

*Represents approximately 6,000,000 speakers. †Approximate.

Table 43: Burmic Languages

	areas where spoken	number of speakers
Burmese-Lolo		
Burmish branch	western China, Myanmar, Indochina	
Burmese and its dialects (Rangoon, Mergui, Intha, Danu, Yaw, Taungyo, Tavoyan, Arakanese)		22,170,000
Maru (Lawng)		
Lashi (Letsi)		
Lolo branch		
Northern group: Lisu, Yi, Lolopho		
Southern group: Hani (Akha), Lahu		
Na-hsi (Naxi)*		
Ch'iang*		
Kachinish		
Kachin (Ching-p'o)†	Kachin and Shan states in Myanmar, adjoining Assam in India, Yunnan in China	600,000
Atsi (Tsaiwa)		
Kukish (Kuki-Chin)		
Kuki branch	border region of Myanmar–southern Assam–Bangladesh	447,000
Central Kuki: Lushai, Lakher, Lai (Haka)		
Northern Kuki: Thado, Kamhau (Tiddim), Siyin (Sizang)		
Southern Kuki: Sho, Yawdwin, Chinbok, Khami (with Khimi)		
Old Kuki: Rangkhol, Bete (Biate); Anai, Lamgang; Aimol, Purum	spoken by dispersed tribes driven out from their original home on the Myanmar–India border	
Western Kuki: Empeo, Maram, Kwoireng, Kabui, Khoirao‡		
Lahupa languages§ (Tangkhul, Maring, Khoibu)		
Nāgā branch	Nāgāland (India), Myanmar	776,000
Northern: Ao		
Eastern group: Rengma, Sema (Simi), Angami		
Lepchā (Rong)‖	Sikkim (India), eastern Nepal, western Bhutan	38,000

*These stand in a not clearly defined relationship to Lolo. †For cultural, nonlinguistic reasons, Maru, Lashi, and Atsi are often grouped with Kachin. ‡In Manipur and Cāchār. §These languages are similar to Kuki. ‖Usually compared to the northern Nāgā branch of Kukish but has Baric and Himalayish affinities.

(such as Bodish, Himalayish, Kirantish, Burmish, Kachinish, Kukish) should be considered as the classificatory peaks around which other Sino-Tibetan languages group themselves as members or more or less distant relatives. Certainly the stage has not yet been reached in which definite boundaries can be laid down and ancestral Proto-, or Common, Tibetic and Proto-, or Common, Burmic can be undisputedly reconstructed.

LANGUAGE AFFILIATIONS

Tibeto-Burman languages. *Tibetic languages.* The Tibetic (also called the Bodic, from Bod, the Tibetan name for Tibet) division comprises the Bodish-Himalayish, Kirantish, and Mirish language groups.

Burmic languages. The Burmic division comprises Burmish, Kachinish, and Kukish.

A number of Tibeto-Burman languages that are difficult to classify have marginal affiliations with Burmic. The Luish languages (Andro, Sengmai, Kadu, Sak, and perhaps also Chairel) in Manipur, India, and adjacent Myanmar resemble Kachin; Nung (including Rawang and Trung) in Kachin state in Myanmar and in Yunnan, China, has similarities with Kachin; and Mikir in Assam, as well as Mru and Meithei in India, Bangladesh, and Myanmar, seem close to Kukish.

Baric languages. The Baric, or Bodo-Garo, division consists of a number of languages spoken in Assam and falls into a Bodo branch (not to be confused with Bodic-Tibetic, and Bodish, a subdivision of Tibetic) and a Garo branch.

Table 44: Baric (Bodo-Garo) Languages

	areas where spoken	number of speakers*
Bodo branch	the plains of Assam	
Bodo		1,000,000
Dimasa		70,000
Garo branch	the hills of Meghālaya	504,000
Achik, Abeng, Dacca Atong, Rabha, Ruga, Koch		

*Approximate.

A group of Sino-Tibetan languages in Nāgāland (Nāgish, not to be confused with the Nāgā branch of Kukish; including Mo Shang, Namsang, and Banpara) has affinities to Baric.

Karenic languages. The Karenic languages of Karen state in Myanmar and adjacent areas in Myanmar and Thailand include the two major languages of the Pho (Pwo) and Sgaw, which have some 3.2 million speakers. Taungthu (Pa-o) is close to Pho, and Palaychi to Sgaw. There are several minor groups.

Chinese, or Sinitic, languages. Chinese as the name of a language is a misnomer. It has been applied to numerous dialects, styles, and languages since the middle of the 2nd millennium BC. Sinitic is a more satisfactory designation for covering all these entities and setting them off from the Tibeto-Karen group of Sino-Tibetan languages. *Han* is a Chinese term for Chinese as opposed to non-Chinese languages spoken in China. The Chinese terms for Modern Standard Chinese are *p'u-t'ung-hua* "common language" and *kuo-yü* "national language" (the latter term is used in Taiwan). Terminology

Reconstructed prehistoric Chinese is known as Proto-Sinitic (or Proto-Chinese); the oldest historic language of China is called Archaic, or Old, Chinese (8th–3rd centuries BC), and that of the next period up to and including the T'ang dynasty (AD 618–907) is known as Ancient, or Middle, Chinese. Languages of later periods include Old, Middle, and Modern Mandarin (the name Mandarin is a translation of *kuan-hua,* "civil servant language"). Through history the Sinitic language area has constantly expanded from the "Middle Kingdom" around the eastern Huang Ho (Yellow River) to its present size. The persistence of a common, nonphonetic writing system for centuries explains why the word dialect rather than language has had widespread usage for referring to the modern speech forms. The present-day spoken languages are not mutually intelligible (some are further apart than Portuguese is from Italian), and neither are the major subdivisions within each group. The variation is slightest in the western and southwestern provinces and greatest along the Huang Ho and in the coastal areas. Table 45 gives the percentage of Chinese people speaking each of the various Chinese languages.

Table 45: Distribution of the Chinese Languages

	areas in China where spoken	ratio of speakers to total population (percentage)*
Mandarin	China, north of the Yangtze River; a narrow belt south of the Yangtze River in Kiangsi, Anhwei, and Kiangsu; Szechwan, Yunnan, and Kweichow; small parts of Hunan	66†
Wu	Kiangsu and Chekiang	7.8
Hsiang	Hunan	4.4
Kan	Kiangsi and a corner of Hupei	2.2
Hakka	east and north Kwangtung; parts of Fukien, Kiangsi, Kwangsi, Taiwan, Hunan, Szechwan; sporadically in other provinces	3.4
Yüeh (Cantonese)	Kwangtung, south Kwangsi, Hong Kong, Macau	4.6
Min§	Fukien, east Kwangtung, Hainan, Taiwan, south Chekiang	3.7

*The 1995 estimate, projected from the 1990 census, is as follows: Mandarin 793,000,000; Wu 94,000,000; Min 45,000,000; Kan 27,000,000; Hakka 41,000,000; Yüeh 55,000,000; Hsiang 53,000,000. †Figure applies to those people speaking Mandarin as the mother tongue as opposed to speaking it as a second language. §Min is subdivided into Northern Min, in north Fukien with a little over 1 percent of the national population, and Southern Min, in the remainder of the Min-speaking area with 2.5 percent of the population.

A vernacular written tradition exists mainly in Peking Mandarin and in Cantonese. An unwritten storytelling tradition has survived in most languages. The school and radio language is Modern Standard Chinese in China as well as in Taiwan and Singapore. In Hong Kong, Cantonese prevails as the language of education and in the communication media, but efforts are now made to adopt Modern Standard Chinese as a norm. The same orthographic system is employed, with some variations, by all speakers of Chinese.

Non-Chinese Sino-Tibetan languages of China include some Lolo-type languages (Burmish)—Yi, with nearly 7 million speakers in Yunnan, Szechwan, Kweichow, and Kwangsi; Hani (Akha) with about 500,000 speakers in Yunnan; Lisu, with approximately 610,000 speakers in Yunnan; Lahu, with about 440,000 speakers in Yunnan; and Na-hsi, with approximately 300,000 speakers mostly in Yunnan and Szechwan. Other Sino-Tibetan languages in Yunnan and Szechwan are Kachin and the closely related Atsi (Tsaiwa); Achang, Nu, Pumi (Primi), Ch'iang, Gyarung, Hsifan; and Pai (Minchia, probably a separate branch within Sinitic).

Linguistic characteristics

COMMON FEATURES

At the end of the 18th and during the first half of the 19th century a great number of languages were investigated by Western scholars in the Himalayas, in India, and in China, and word lists and grammatical sketches began to appear. By the late 19th century a foundation had been laid for Sino-Tibetan comparative studies.

The comparative method for determining genetic relationship among languages was worked out in detail for Indo-European during the latter part of the 19th century. It rests on the assumption that sound correspondences in related words and morphological units, as well as structural similarities on all levels (phonology, morphology, syntax), can be explained in terms of a reconstructed common language, or protolanguage. Structural or typological similarities, however, are in many cases due to interaction among contiguous languages over a long time, creating so-called linguistic, or language, areas. The morphology and syntax of the Sino-Tibetan languages are for the most part rather simple and nonspecific, and the length of time The comparative method involved in the separation of subfamilies and divisions is such that comparative phonological statements are often difficult to reduce to concise correspondences and laws.

A number of features have been delineated as common for the Sino-Tibetan languages. Many of them can be shown to be of a typological nature, the result of diffusion and underlying unrelated language strata.

Typological similarities. *Monosyllabicity.* The vast majority of all words in all Sino-Tibetan languages are of one syllable, and the exceptions appear to be secondary (*i.e.,* words that were introduced at a later date than Common, or Proto-, Sino-Tibetan). Some suffixes in Tibeto-Burman are syllabic, thus adding a syllable to a word, but they have a highly reduced set of vowels and tones ("minor syllables"). These features are, however, shared by contiguous languages (namely, those of Austroasiatic stock and Miao-Yao) and are not clearly attributable to Sino-Tibetan on the basis of shared basic vocabulary items.

Tonality. Most Sino-Tibetan languages possess phonemic tones, which indicate a difference in meaning in otherwise similar words. There are no tones in Purik, a Western Tibetan language; Ambo, a Northern Tibetan tongue; and Newari of Nepal. Balti, another Western Tibetan language, has pitch differences in polysyllabic nouns. The tones of the remaining Tibetan dialects can be accounted for by positing an original and older system of voiced and voiceless initial sounds that eventually resulted in tones. In several Himalayish languages, tones are linked with articulatory features connected with the end of the syllable or are linked with stress features, as also in Kukish Lepchā (Rong).

Most Baric languages lack tones altogether; and Burmic, Karenic, and Sinitic tonal systems can be reduced to two basic tones ultimately probably accounted for by different syllabic endings. What can be reconstructed for Proto-Sino-Tibetan, the language from which all the modern Sino-Tibetan languages developed, are a set of conditioning factors (as, for example, certain syllabic endings) that resulted in tones; the tones themselves cannot be reconstructed. Again the features that encouraged the development of tones are not uniquely Sino-Tibetan; similar conditions have produced similar effects in Tai and Miao-Yao and—within the Austroasiatic languages—in Vietnamese and in the embryonic form of two registers (pitches or vocal qualities) also in Cambodian.

Prefixes and suffixes

Affixation. Most Sino-Tibetan languages possess or can be shown to have at one time possessed derivational and morphological affixes—*i.e.,* word elements attached before or after or within the main stem of a word that change or modify the meaning in some way. Many prefixes can be reconstructed for Proto-Sino-Tibetan: *s-* (causative), *m-* (intransitive), *b-*, *d-*, *g-*, and *r-*, and many more for certain language divisions and units. Among the suffixes, *-s* (used with several types of verbs and nouns), *-t*, and *-n* are inherited from the protolanguage. The problem of whether Proto-Sino-Tibetan made use of *-r-* and *-l-* infixes (besides perhaps semivocalic infixes) has not been solved. Whether clusters containing these sounds were the result of prefixation to roots beginning in *r* and *l* (and *y*) or came about through infixation is not clear.

Initial consonant alternation. Voiced and voiceless initial stops alternate in the same root in many Sino-Tibetan languages, including Chinese, Burmese, and Tibetan (voiced in intransitive, voiceless in transitive verbs). The German Oriental scholar August Conrady linked this morphological system to the causative *s-* prefix, which was supposed to have caused devoicing of voiced stops. (Voicing is the vibration of the vocal cords, as occurs, for example, in the sounds *b, d, g, z,* and so on. Devoicing, or voicelessness, is the pronunciation of sounds without vibration of the vocal cords, as in *p, t, k, s.*) Such alternating of the initial consonant cannot itself be reconstructed for the protolanguage.

Vowel alternation. The morphological use of vowel gradation (called ablaut) is well known from Indo-European languages (*e.g.,* the vowel change in English *sing, sang, sung*) and is found in several Sino-Tibetan languages, including Chinese and Tibetan. In Tibetan the various forms of the verbs are differentiated in part by vowel alternation; in Sinitic some related words (known as word families) are kept apart by vowel alternation. Some conditioning factor outside the vowel (perhaps stress or sandhi, the modification of a sound according to the surrounding sounds) may have been responsible for the Sino-Tibetan ablaut systems.

Indistinct word classes. Especially in the older stages of Sino-Tibetan, the distinction of verbs and nouns appears blurred; both overlap extensively in the Old Chinese writing system. Philological tradition as well as Sinitic reconstruction show, however, that frequently, when the verb and the noun were written alike, they were pronounced differently, the difference manifesting itself later in the tonal system. Verbs and nouns also used different sets of particles.

Classifiers for nouns

Use of noun classifiers. The Sino-Tibetan noun is typically a collective term, designating all members of its class, like the English *man* used to signify "all human beings." In a number of modern Sino-Tibetan languages, such a noun can be counted or modified by a demonstrative pronoun only indirectly through a smaller number of noncollective nouns, called classifiers, in constructions such as "one *person* man," "one *animal* dog," and so on, much like parallel cases in Indo-European (in English, "one head of cattle"; in German, *ein Kopf Salat* "one head of lettuce"). The phenomenon is absent in Tibetan and appears late in Burmese and Chinese. Furthermore, classifiers are not exclusively Sino-Tibetan; they exist also in Miao-Yao, Tai, Austric, and Japanese. In Classical Chinese, Tai, and Burmese, the classifier construction follows the noun; in modern Chinese, as in Miao-Yao, it precedes it. Classifiers are of later origin and do not belong to Proto-Sino-Tibetan.

Word order. Although the word order of subject–object–verb and modified–modifier prevails in Tibeto-Burman, the order subject–verb–object and modifier–modified occurs in Karenic. In this respect Chinese is like Karen, although Old Chinese shows remnants of the Tibeto-Burman word order. Tai employs still another order: subject–verb–object, and modified–modifier, like Austric but unlike Miao-Yao, which follows the Karen and Chinese model. Word order, even more than any of the other distinguishing features, points to diffusion from several centres, or to unrelated substrata.

Phonological correspondences. The hypothesis that the Sino-Tibetan languages are all related and derive from a common source depends on phonological correspondences in shared vocabulary more than on any other argument. It is ironic that the clearest and most convincing results should have been obtained from studies of the Sinitic-Tai similarities, which probably do not indicate a true case of genetic relationship. In 1942 most of the words in this grouping were shown to be cultural loans (then thought of as Chinese loanwords in Tai, now believed to a very large extent to be borrowings in the opposite direction).

A comparison of Old Chinese and Old Tibetan made by Walter Simon in 1929, although limited in some ways, pointed to enough sound resemblances in important items of basic vocabulary to eliminate the possibility of coincidental similarities between unrelated languages. A few examples of similar words in Old Tibetan and Old Chinese, respectively, follow: "bent," *gug* and *gyuk;* "eye," *myig* and *myəkw;* "friend," *grogs* and *gyəgw;* "kill," *gsod* and *sriat;* "onion," *btsong* and *tshung;* "rise," *lang* and *rang;* "single, one," *gcig* and *tyik;* "sun," *nyi* and *nyit.* The American linguist Paul Benedict brought in material from other Sino-Tibetan languages and laid down the rule that the comparative linguist should accept perfect phonetic correspondences with inexact though close semantic equivalences in preference to perfect semantic equivalences with questionable phonetic correspondences. New material and competent descriptions later made it possible to reconstruct important features of common ancestral languages within major divisions of Sino-Tibetan (notably Lolo, Baric, Tibetic, Kachin, Kukish, Karenic, Sinitic).

INTERRELATIONSHIP OF THE LANGUAGE GROUPS

The position of Proto-Sino-Tibetan can be defined in terms of a chain of interrelated languages and language groups:

Chain of interrelated languages

Sinitic is connected with Tibetic through a body of shared vocabulary and typological features, similarly Tibetic with Baric, Baric with Burmic, and Burmic with Karenic. The chain continues at both ends, connecting Sinitic to Tai and Tai to Austronesian and also connecting Karenic with Austroasiatic. Considerations of basic vocabulary versus cultural loans and diffusion versus inheritance have led scholars to believe that only the members of the chain from Sinitic to Karenic share a common ancestral language; especially Sinitic and Karenic are under suspicion for containing only superstrata of Sino-Tibetan origin.

Proto-Tibeto-Burman. The Proto-Tibeto-Burman language was monosyllabic. Some grammatical units may have had the form of minor syllables before the major syllable (**ma-*, **ba-*) or after the major syllable (**-ma*, **-ba*). (An asterisk [*] indicates that the form it precedes is unattested and has been reconstructed as a possible ancestral form.) The consonants were three voiceless stops (*p*, *t*, *k*), which were aspirated in absolute initial position, three voiced stops (*b*, *d*, *g*), and three nasal sounds (*m*, *n*, *ŋ* [as the *-ng* in "sing"]). There were five continuant sounds (*s*, *z*, *r*, *l*, and *h*) and two semivowels (*w*, *y*). In final position there was only one set of stops, but there were a number of initial and final clusters mainly resulting from the addition of prefixes and suffixes. Three degrees of vowel opening existed with two members in each: *i* and *u*, *e* and *o*, *a* and *aa* (short and long *a*). Length may have been relevant also with the *i* and *u* and *e* and *o* vowels. The conditioning factors that led to the development of tones can be shown to have been voiced–voiceless contrast in initial and final consonants and consonant clusters. Because the conditioning factors were involved with morphological process (affixation and consonant alternation), tonal systems could also acquire certain grammatical or structural functions. An independent morphological system involved or resulted in vowel alternation (see also above).

The sound system of Proto-Karenic appears closely related to that of Proto-Tibeto-Burman. The tonal classes can be reduced to two, which connect Karen to Burmic, Sinitic, Tai, and Miao-Yao.

Proto-Sinitic. Greater dissimilarity is encountered with respect to Proto-Sinitic. The contrast of aspirated and unaspirated voiceless stops in initial position is most likely the result of lost initial cluster elements as in Proto-Tibeto-Burman. The voiced stops possibly also had the aspirated–unaspirated distinction. Unlike Tibeto-Burman, two series of stops in syllable final position are posited for Old Chinese, but it is not clear if the contrast involved voicing or other features. One series is in general without an exact correspondence in Tibeto-Burman languages, but Burmish Maru has final stops in a number of these words. Similar isolated cases are found in Tibetan and in Tai.

Old Chinese has two more relevant points of articulation, or sound-producing positions of the mouth, than Proto-Tibeto-Burman: palatal (in which the tongue blade touches the palate) and retroflex (in which the tip of the tongue is curled upward toward the palate). But these two types of sounds may be explained as the result of influence from lost Proto-Sinitic medial sounds (a palatal *-y-* and a retroflex *-r-*). The relationship between these specific medial sounds and similar elements in Tibeto-Karen is, however, not certain. Dental affricate sounds in Old Chinese, which begin as stops with complete stoppage of the breath stream and conclude as fricatives with incomplete air stoppage and audible friction, can at least be explained partly as metathesized (transposed) forms of prefix *s-* plus a dental sound in Proto-Sinitic (*e.g.*, *st* changes to *ts*). Old Chinese possessed initial consonant clusters containing *-l-* as a second element, so Proto-Sinitic can reasonably be supposed to have had the same three medial elements as Proto-Tibeto-Burman: *-y-*, *-l-*, and *-r-*. There are few, if any, traces in Old Chinese of the more complicated clusters and the minor syllables of Tibeto-Burman.

Vowel system of Old Chinese

The vowel system of Old Chinese as reconstructed (1940) by the linguist Bernhard Karlgren to account especially for the language of the *Shih ching*, an anthology of Chinese poetry compiled in the 6th–5th centuries BC, seems surprisingly complicated as compared to that of Proto-Tibeto-Burman. Probably some of the vowels can be explained as diphthongs or as combinations of vowels plus specific classes of consonants (*e.g.*, labialized, retroflex, palatal).

As in Karen and Burmese-Loloish, the tones of Sinitic can be reduced to two (with some syllables or syllabic types being neutral or unaffected by the tonal system; the modern Sinitic languages have from two to as many as eight or nine tones). Monosyllabicity of roots and morphological affixation were characteristic features of Proto-Sino-Tibetan as they were of Proto-Tibeto-Karen.

Languages of the family

CHARACTERISTICS OF THE MODERN CHINESE LANGUAGES

All modern Sinitic languages—*i.e.*, the "Chinese dialects"—share a number of important typological features. They have a maximum syllabic structure of the type consonant–semivowel–vowel–semivowel–consonant. Some languages lack one set of semivowels, and, in some, gemination (doubling) or clustering of vowels occurs. The languages also employ a system of tones (pitch and contour), with or without concomitant glottal features, and occasionally stress. For the most part, tones are lexical (*i.e.*, they distinguish otherwise similar words); in some languages tones also carry grammatical meaning. Nontonal grammatical units (*i.e.*, affixes) may be smaller than syllables, but usually the meaningful units consist of one or more syllables. Words can consist of one syllable, of two or more syllables each carrying an element of meaning, or of two or more syllables that individually carry no meaning. For example, Modern Standard Chinese *t'ien* "sky, heaven, day" is a one-syllable word; *jih-t'ou* "sun" is composed of *jih* "sun, day," a word element that cannot occur alone as a word, and the noun suffix *t'ou*; and *hu-t'ier* "butterfly" consists of two syllables, each having no meaning in itself (this is a rare type of word formation). The Southern languages have more monosyllabic words and word elements than the Northern ones.

Sinitic grammatical characteristics

The Sinitic languages distinguish nouns and verbs with some overlapping, as do Sino-Tibetan languages in general. There are noun suffixes that form different kinds of nouns (concrete nouns, diminutives, abstract nouns, and so on), particles placed after nouns indicating relationships in time and space, and verb particles for modes and aspects. Adjectives act as one of several kinds of verbs. Verbs can occur in a series (concatenation) with irreversible order (*e.g.*, the verbs "take" and "come" placed next to one another denote the concept "bring"). Nouns are collective in nature, and only classifiers (see above) can be counted and referred to singly. Specific particles are used to indicate the relationship of nominals (*e.g.*, nouns and noun phrases) to verbs, such as transitive verb–object, agent–passive verb; in some of the languages this system forms a sentence construction called ergative, in which all nominals are marked for their function and the verb stays unchanged. Final sentence particles convey a variety of meanings (defining either the whole sentence or the predicate), such as "question, command, surprise, new situation." The general word order of subject–verb–object and complement and modifier–modified is the same in all the languages, but the use of the preposed particles and verbs in a series varies considerably. Grammatical elements of equal or closely related values in various languages are very often not related in sounds.

The Sinitic languages fall into a Northern and a Southern group. The Northern languages (Mandarin dialects) are more similar to each other than are the Southern (Wu, Hsiang, Kan, Hakka, Yüeh, Min).

Modern Standard Chinese (Mandarin). Modern Standard Chinese is based on the Peking (Beijing) dialect, which is of the Northern, or Mandarin, type. It employs about 1,300 different syllables. There are 22 initial consonants, including stops (made with momentary, complete closure in the vocal tract), affricates (beginning as stops but ending with incomplete closure), aspirated consonants, nasals, fricatives, liquid sounds (*l*, *r*), and a glottal stop. The medial semivowels are *y* (*i*), *ɥ* (*ü*), and *w* (*u*). In final position, the following occur: nasal consonants, *ṛ* (retroflex r), the semivowels *y* and *w*, and the combinations *ŋr* (nasalization plus *r*) and *wr* (rounding plus *r*). There are nine vowel sounds, including three varieties of *i* (retroflex, apical, and palatal). Several vowels combine into clusters.

There are four tones: (1) high level, (2) high rising crescendo, (3) low falling diminuendo with glottal friction (with an extra rise from low to high when final), and (4) falling diminuendo. Unstressed syllables have a neutral tone, which depends on its surroundings for pitch. Tones in sequences of syllables that belong together lexically and syntactically ("sandhi groups") may undergo changes known as tonal sandhi, the most important of which causes a third tone before another third tone to be pronounced as a second tone. The tones influence some vowels (notably *e* and *o*), which are pronounced more open in third and fourth tones than in first and second tones.

A surprisingly low number of the possible combinations of all the consonantal, vocalic, and tonal sounds are utilized. The vowels *i* and *ü* and the semivowels *y* and *ɥ* never occur after velar sounds (*e.g.*, *k*) and occur only after the palatalized affricate and sibilant sounds (*e.g.*, *tś*), which in turn occur with no other vowels and semivowels.

Many alternative interpretations of the distinctive sounds of Chinese have been proposed; the interaction of consonants, vowels, semivowels, and tones sets Modern Standard Chinese apart from many other Sinitic languages and dialects and gives it a unique character among the major languages of the world. The two most widely used transcription systems (romanizations) are Wade-Giles (first propounded by Sir Thomas Francis Wade in 1859 and later modified by Herbert A. Giles) and the official Chinese transcription system today, known as the *pinyin zimu* ("phonetic spelling") or simply Pinyin (adopted in 1958). In Wade-Giles, aspiration is marked by ' (*p'*, *t'*, and so on). The semivowels are *y*, *yü*, and *w* in initial position; *i*, *ü*, and *u* in medial; and *i* and *u* (but *o* after *a*) in final position. Final retroflex *r* is written *rh*. The tones are indicated by raised figures after the syllables (^{1}, ^{2}, ^{3}, ^{4}).

Transcription systems

The Pinyin system indicates unaspirated stops and affricates by means of traditionally voiced consonants (*e.g.*, *b*, *d*) and aspirated consonants by voiceless sounds (*e.g.*, *p*, *t*). The semivowels are *y*, *yu*, and *w* initially; *i*, *ü*, and *u* medially; and *i* and *u* (*o* after *a*) finally. Final retroflex *r* is written *r*. The tones are indicated by accent markers, 1 = ˉ, 2 = ´, 3 = ˇ, 4 = ` (*e.g.*, *mā*, *má*, *mǎ*, *mà* = Wade-Giles *ma*^{1}, *ma*^{2}, *ma*^{3}, *ma*^{4}).

Wade-Giles is used in the following discussion of Modern Standard Chinese grammar.

The most common suffixes that indicate nouns are *-erh* (as in *ch'ang-erh* "song"; compare *ch'ang* "sing"), *-tzu* (as in *fang-tzu* "house"), and *-t'ou* (as in *mu-t'ou* "wood"). A set of postposed noun particles express space and time relationships (*-li* "inside," *-hou* "after"). An example of a verbal affix is *-chien* in *k'an-chien* "see" and *t'ing-chien* "hear." Important verb particles are *-le* (completed action), *-kuo* (past action), *-chih* or *-che* (action in progress). The directional verbal particles *-lai* "toward speaker" and *-ch'ü* "away from speaker" and some verbal suffixes can be combined with the potential particles *te* "can" and *pu* "cannot"—*e.g.*, *na-ch'u-lai* "take out," *na-pu-ch'u-lai* "cannot take out"; *t'ing-chien* "hear," *t'ing-te-chien* "can hear." The particle *te* indicates subordination and also gives nominal value to forms for other parts of speech (*e.g.*, *wo* "I," *wo-te* "mine," *wo-te shu* "my book," *lai* "to come," *lai-te* "coming," *lai-te jen* "a person who comes"). The most important sentence particle is *le*, indicating "new situation" (*e.g.*, *hsia-yü-le* "now it is raining," *pu-lai-le* "now there is no longer any chance that he will be coming"). *Ko* is the most common noun classifier (*i* "one," *i-ko-jen* "one person"); others are *so* (*i-so-fang-tzu* "one house") and *pen* (*liang-pen-shu* "two books").

Adjectives can be defined as qualitative verbs (*hao* "to be good") or stative verbs (*ping* "to be sick"). There are equational sentences with the word order subject–predicate—*e.g.*, *wo-shih Pei-ching-jen* "I am a Peking-person (*i.e.*, a native of Peking)"—and narrative sentences with the word order subject (or topic)–verb–object (or complement)—*e.g.*, *wo ch'ih-fan* "I eat rice," *wo chu tsai Pei-ching* "I live in Peking." The preposed object takes the particle *pa* (*wo ta t'a* "I beat him," *wo pa t'a ta-le i-tun* "I gave him a beating"), and the agent of a passive construction takes *pei* (*wo pei t'a ta-le i-tun* "I was given a beating by him").

Standard Cantonese. The most important representative of the Yüeh languages is Standard Cantonese of Canton, Hong Kong, and Macau. It has fewer initial consonants than Modern Standard Chinese (*p*, *t*, *ts*, *k* and the corresponding aspirated sounds *ph*, *th*, *tsh*, *kh*; *m*, *n*, *ŋ*; *f*, *s*, *h*; *l*, *y*), only one medial semivowel (*w*), more vowels than Modern Standard Chinese, six final consonants (*p*, *t*, *k*, *m*, *n*, *ŋ*), and two final semivowels (*y* and *w*). The nasals *m* and *ŋ* occur as syllables without a vowel.

There are three tones (high, mid, low) in syllables ending in *-p*, *-t*, and *-k*; six tones occur in other types of syllables (mid level, low level, high falling, low falling, high rising, low rising). Two tones are used to modify the meaning of words (high level °, and low-to-high rising *), as in *yin*° "tobacco" from *yin* "smoke," and *nöy** "daughter" from *nöy* "woman." Some special grammatical words also have the tone °. There is no neutral tone and little tonal sandhi.

Cantonese tonal system

There are more than 2,200 different syllables in Standard Cantonese, or almost twice as many as in Modern Standard Chinese. The word classes are the same as in Modern Standard Chinese. The grammatical words, although phonetically unrelated, generally have the same semantic value (*e.g.*, the subordinating and nominalizing particle *kɛ*, Modern Standard Chinese *te*; *m* "not," Modern Standard Chinese *pu*; the verbal particle for "completed action" and the sentence particle for "new situation," both *le* in Modern Standard Chinese, are Standard Cantonese *tsɔ* and *lɔ*, respectively). A classifier preceding a noun in subject position (before the verb) functions as a definite article (*e.g.*, *tsek sün* "the boat").

Min languages. The most important Min language is Amoy from the Southern branch of Min. The initial consonants are the same as in Standard Cantonese with the addition of two voiced stops (*b* and *d*) and one voiced affricate (*dz*), developed from original nasals. There are two semivowels (*y*, *w*), six vowels and several vowel clusters, plus the syllabic nasal sounds *m* and *ŋ* functioning as vowels, the same finals as in Standard Cantonese, and, in addition, a glottal stop (ˀ) and a meaning-bearing feature of nasalization, as well as a combination of the last two features. There are two tones in syllables ending in a stop, five in other syllables. Tonal sandhi operates in many combinations.

Fuchow is the most important language of the Northern branch of Min. The very extensive sandhi affects not only tones but also consonants and vowels, so that the phonetic manifestation of a syllable depends entirely on interaction with the surroundings. There are three initial labial sounds (*p*, *ph*, *m*), five dental sounds (*t*, *th*, *s*, *l*, *n*), three palatal sounds (*tś*, *tśh*, *ń*), and five velars (*k*, *kh*, *h*, ˀ, and *ŋ*). Syllables can end in *-k*, *-ŋ*, ˀ (glottal stop), a semivowel, or a vowel. The tones fall into two classes: a comparatively high class comprising high, mid, high falling, and high rising (only in sandhi forms) and a rather low one, comprising low rising and low rising-falling (circumflex). Certain vowels and diphthongs occur only with the high class, others occur only with the low class, and the vowel *a* occurs with both classes. Sandhi rules can cause tone to change from low class to high class, in which case the vowel also changes.

Other Sinitic languages. *Hakka.* Of the different Hakka dialects, Hakka of Mei-chou (formerly Mei-hsien) in Kwangtung is best known. It has the same initial consonants, final consonants, and syllabic nasals as Standard Cantonese; the vowels are similar to those of Modern Standard Chinese. Medial and final semivowels are *y* and *w*. There are two tones in syllables with final stops, four in the other syllabic types.

Süchow. Süchow is usually quoted as representative of the Wu languages. It is rich in initial consonants, with a contrast of voiced and voiceless stops as well as palatalized and nonpalatalized dental affricates, making 26 consonants in all. (Palatalized sounds are formed from nonpalatal sounds by simultaneous movement of the tongue toward the hard palate. Dental affricates are sounds produced with the tongue tip at first touching the teeth and then drawing slightly away to allow air to pass through, producing a hissing sound.) Medial semivowels

Characteristics of Wu languages

are as in Modern Standard Chinese. In addition, there are also 10 vowels and 4 syllabic consonants (*l, m, n, ŋ*); *-n* and *-ŋ* occur in final position, as do the glottal stop and nasalization.

Shanghai dialect. The Shanghai dialect belongs to Wu. The use of only two tones or registers (high and low) is prevalent; these are related in an automatic way to the initial consonant type (voiceless and voiced).

Hsiang languages. The Hsiang languages, spoken only in Hunan, are divided into New Hsiang, which is under heavy influence from Mandarin and includes the language of the capital Ch'ang-sha, and Old Hsiang, more similar to the Wu languages, as spoken for instance in Shuang-feng. Old Hsiang has 28 initial consonants, the highest number for any major Sinitic language, and 11 vowels, plus the syllabic consonants *m* and *n.* It also uses five tones, final *-n* and *-ŋ*, and nasalization, but no final stops.

HISTORICAL SURVEY OF CHINESE

The early contacts. Old Chinese vocabulary already contained many words not generally occurring in the other Sino-Tibetan languages. The words for "honey" and "lion," probably also "horse," "dog," and "goose," are connected with Indo-European and were acquired through trade and early contacts. (The nearest known Indo-European languages were Tocharian and Sogdian.) A number of words have Austroasiatic cognates and point to early contacts with the ancestral language of Muong-Vietnamese and Mon-Khmer—*e.g.*, the name of the Yangtze River, **kruŋ*, is still the word for "river"—Cantonese *kɔŋ*, Modern Standard Chinese *chiang*, pronounced *kroŋ* and *kloŋ* in some modern Mon-Khmer languages. Words for "tiger," "ivory," and "crossbow" are also Austroasiatic. The names of the key terms of the Chinese calendar ("the branches") have this same non-Chinese origin. It has been suggested that a great many cultural words that are shared by Chinese and Tai are Chinese loanwords from Tai. Clearly, the Chinese received many aspects of culture and many concepts from the Austroasiatic and Austro-Tai peoples whom they gradually conquered and absorbed or expelled.

From the 1st century AD, China's contacts with India, especially through the adoption of Buddhism, led to Chinese borrowing from Indo-Aryan (Indic) languages, but, very early, native Chinese equivalents were invented. Sinitic languages have been remarkably resistant to direct borrowing of foreign words. In modern times this has led to an enormous increase in Chinese vocabulary without a corresponding increase in basic meaningful syllables. For instance, *t'ieh-lu* "railroad" is based on the same concept expressed in the French *chemin de fer,* using *t'ieh* "iron" and *lu* "road"; likewise, *tien-hua* "telephone" is a compound of *tien* "lightning, electricity" and *hua* "speech." A number of such words were coined first in Japanese by means of Chinese elements and then borrowed back into Chinese. The reason that China has avoided the incorporation of foreign words is first and foremost a phonetic one; such words fit very badly into the Chinese pattern of pronunciation. A contributing factor has been the Chinese script, which is ill-adapted to the process of phonetic loans. In creating new words for new ideas, the characters have sometimes been determined first and forms have arisen that cannot be spoken without ambiguity ("sulfur" and "lutecium" coalesced as *liu,* "nitrogen" and "tantalum" as *tan*). It is characteristic of Modern Standard Chinese that the language from which it most freely borrows is one from its own past: Classical Chinese. In recent years it has borrowed from Southern Sinitic languages under the influence of statesmen and revolutionaries (Chiang Kai-shek was originally a Wu speaker and Mao Zedong a Hsiang speaker). Influence from English and Russian (in word formation and syntax) has been increasingly felt.

Chinese practices in word borrowing and word creation

Pre-Classical Chinese. The history of the Chinese language can be divided into three periods, pre-Classical (*c.* 1500 BC–*c.* AD 200), Classical (*c.* 200–*c.* 1920), and post-Classical Chinese (with important forerunners as far back as the T'ang dynasty).

The pre-Classical Chinese is further divided into Oracular Chinese (Shang dynasty [18th–12th centuries BC]), Archaic Chinese (Chou and Ch'in dynasties [1111–206 BC]), and Han Chinese (Han dynasty [206 BC–AD 220]).

Oracular Chinese is known only from rather brief oracle inscriptions on bones and tortoise shells. Archaic Chinese falls into Early, Middle (*c.* 800–*c.* 400 BC), and Late Archaic. Early Archaic is represented by bronze inscriptions, parts of the *Shu ching* ("Classic of History"), and parts of the *Shih ching* ("Classic of Poetry"). From this period on, many important features of the pronunciation of the Chinese characters have been reconstructed (see below). The grammar depended to a certain extent on unwritten affixes. The writing system kept apart forms with or without medial consonants, which in some cases were meaningful infixes. Early Archaic Chinese possessed a third-person personal pronoun in three cases (nominative and genitive *gyəg,* accusative *tyəg,* and another special genitive *kywat,* used only with concepts intimately connected with the owner). No other kind of written Chinese until the post-Classical period possessed a nominative of the third-person pronoun, but the old form survived in Cantonese (*khöy*) and is probably also found in Tai (Modern Thai *khǎw*).

Middle Archaic Chinese is the language of some of the earliest writings of the Confucian school. Important linguistic changes that had occurred between the Early and Middle phases became still more pronounced in Late Archaic, the language of the two major Confucian and Taoist writers, Mencius (Meng-tzu) and Chuang-tzu, as well as of other important philosophers. The grammar by then had become more explicit in the writing system, with a number of well-defined grammatical particles, and it can also be assumed that the use of grammatical affixes had similarly declined. The process used in verb formation and verb inflection that later appeared as tonal differences may at this stage have been manifested as final consonants or as suprasegmental features, such as different types of laryngeal phonation. The word classes included nouns, verbs, and pronouns (each with several subclasses), and particles. The use of a consistent system of grammatical particles to form noun modifiers, verb modifiers, and several types of embedded sentences (*i.e.,* sentences that are made to become parts of another independent sentence) became blurred in Han Chinese and was gone from written Chinese until the emergence of post-Classical Chinese. In Modern Standard Chinese the subordinating particle *te* combines the functions of several Late Archaic Chinese particles, and the verb particle *le* and the homophonous sentence particle *le* have taken over for other Late Archaic forms.

Evolution of Chinese grammar

Han and Classical Chinese. Han Chinese developed more polysyllabic words and more specific verbal and nominal (noun) categories of words. Most traces of verb formation and verb conjugation began to disappear. An independent Southern tradition (on the Yangtze River), simultaneous with Late Archaic Chinese, developed a special style, used in the poetry *Ch'u tz'u* ("Elegies of Ch'u"), which was the main source for the refined *fu* (prose poetry). Late Han Chinese developed into Classical Chinese, which as a written idiom underwent few changes during the long span of time it was used. It was an artificial construct, which for different styles and occasions borrowed freely and heavily from any period of pre-Classical Chinese but in numerous cases without real understanding for the meaning and function of the words borrowed.

At the same time the spoken language changed continually, as did the conventions for pronouncing the written characters. Soon Classical Chinese made little sense when read aloud. It depended heavily on fixed word order and on rhythmical and parallel passages. It has sometimes been denied the status of a real language, but it was certainly one of the most successful means of communication in human history. It was the medium in which the poets Li Po (701–762) and Tu Fu (712–770) and the prose writer Han Yü (768–824) created some of the greatest masterpieces of all times and was the language of the Neo-Confucianist philosophy (especially of Chu Hsi [1130–1200]), which was to influence the West deeply. Classical Chinese was also the language in which the Italian Jesuit missionary Matteo Ricci (1552–1610) wrote in his attempt to convert the Chinese empire to Christianity.

Post-Classical Chinese. Post-Classical Chinese, based on dialects very similar to the language now spoken in North China, probably owes its origin to the Buddhist storytelling tradition; the tales appeared in translations from Sanskrit during the T'ang dynasty (618–907). During the Sung dynasty (960–1279) this vernacular language was used for polemic writings; it also appeared in indigenous Chinese novels based on popular storytelling. During and after the Yüan dynasty (1206–1368) the vernacular was used also in the theatre.

Modern Standard Chinese has a threefold origin: the written post-Classical language, the spoken standard of Imperial times (Mandarin), and the vernacular language of Peking. These idioms were clearly related originally, and combining them for the purpose of creating a practical national language was a task that largely solved itself once the signal had been given. The term National Language (*kuo-yü*) had been borrowed from Japanese at the beginning of the 20th century, and, from 1915, various committees considered the practical implications of promoting it. The deciding event was the action of the May Fourth Movement of 1919; at the instigation of the liberal savant Hu Shih, Classical Chinese (also known as *wen-yen*) was rejected as the standard written language. (Hu Shih also led the vernacular literature movement of 1917; his program for literary reform appeared on Jan. 1, 1917.) The new written idiom has gained ground faster in literature than in science, but there can be no doubt that the days of Classical Chinese as a living medium are numbered. After the establishment of the People's Republic of China, some government regulation was applied successfully, and the tremendous task of making Modern Standard Chinese understood throughout China was effectively undertaken. In what must have been the largest-scale linguistic plan in history, untold millions of Chinese, whose mother tongues were divergent Mandarin or non-Mandarin languages or non-Chinese languages, learned to speak and understand the National Language; with this effort, literacy was imparted to great numbers of people in all age groups.

Adoption of the National Language

The writing system. The Chinese writing system is nonalphabetic. It applies a specific character to write each meaningful syllable or each nonmeaningful syllabic that is part of a polysyllabic word.

Pre-Classical characters. When the Chinese script first appeared, as used for writing Oracular Chinese (from *c.* 1500 BC), it must already have undergone considerable development. Although many of the characters can be recognized as originally depicting some object, many are no longer recognizable. The characters did not indicate the object in a primitive nonlinguistic way but only represented a specific word of the Chinese language (*e.g.*, a picture of the phallic altar to the earth is used only to write the word *earth*). It is therefore misleading to characterize the Chinese script as pictographic or ideographic; nor is it truly syllabic, for syllables that sound alike but have different meanings are written differently. Logographic (*i.e.*, marked by a letter, symbol, or sign used to represent an entire word) is the term that best describes the nature of the Chinese writing system.

Verbs and nouns are written by what are or were formerly pictures, often consisting of several elements (*e.g.*, the character for "to love" depicts a woman and a child; the character for "beautiful" is a picture of a man with a huge headdress with ram's horns on top). The exact meaning of the word is rarely deducible from even a clearly recognizable picture, because the connotations are either too broad or too narrow for the word's precise meaning. For example, the picture "relationship of mother to child" includes more facets than "love," a concept that, of course, is not restricted to the mother-child relation, and a man adorned with ram's horns undoubtedly had other functions than that of being handsome to look at, whereas the concept "beautiful" is applicable also to men in other situations, as well as to women. Abstract nouns are indicated by means of concrete associations. The character for "peace, tranquility" consists of a somewhat stylized form of the elements "roof," "heart," and "(wine) cup." Abstract symbols have been used to indicate numbers and local relationships.

Abstract nouns

Related words with similar pronunciations were usually written by one and the same character (the character for "to love, to consider someone good" is a derivative of a similarly written word "to be good"). This gave rise to the most important invention in the development of the Chinese script—that of writing a word by means of another one with the same or similar pronunciation. A picture of a carpenter's square was primarily used for writing "work, craftsman; to work" and was pronounced *kuŋ*; secondarily it was used to write *kuŋ-* (the hyphen stands for an element that was perhaps *s*) "to present," *guŋ* "red," *kuŋ* "rainbow," *kruŋ* "river." During the Archaic period this practice was developed to such a degree that too many words came to be written as one and the same character. In imitation of the characters that already consisted of several components an element was added for each meaning of a character to distinguish words from each other. Thus "red" was no longer written with a single component but acquired an additional component that added the element "silk" on the left; "river" acquired an additional component of "water." The original part of the character is referred to as its phonetic and the added element as its radical.

Ch'in dynasty standardization. During the Ch'in dynasty (221–206 BC) the first government standardization of the characters took place, carried out by the statesman Li Ssu. A new, somewhat formalized style known as seals was introduced—a form that generally has survived until now, with only such minor modifications as were necessitated by the introduction of the writing brush about the beginning of the 1st century AD and printing about AD 600. As times progressed, other styles of writing appeared, such as the regular handwritten form *k'ai* (as opposed to the formal or scribe style *li*), the running hand *hsing*, and the cursive hand *ts'ao*, all of which in their various degrees of blurredness are explicable only in terms of the seal characters.

The Ch'in dynasty standardization comprised more than 3,000 characters. In addition to archaeological finds, the most important source for the early history of Chinese characters is the huge dictionary *Shuo-wen chieh-tzu*, compiled by Hsü Shen about AD 100. This work contains 9,353 characters, a number that certainly exceeds that which it was or ever became necessary to know offhand. Still, a great proliferation of characters took place at special times and for special purposes. The *Kuang-yün* dictionary of 1007 had 26,194 characters (representing 3,877 different syllables in pronunciation). The *K'ang-hsi tzu-tien*, a dictionary of 1716, contains 40,545 characters, of which, however, fewer than one-fourth were in actual use at the time. The number of absolutely necessary characters has probably never been much more than 4,000–5,000 and is today estimated at fewer than that.

Dictionaries of characters

The 20th century. By the 20th century the feeling had become very strong that the script was too cumbersome and an impediment to progress. The desire to obtain a new writing system necessarily worked hand in hand with the growing wish to develop a written language that in grammar and vocabulary approached modern spoken Chinese. If a phonetic writing were to be introduced, the classical language could not be used at all because it deviates so markedly from the modern language. None of the earlier attempts gained any following, but in 1919 a system of phonetic letters (inspired by the Japanese syllabaries called *kana*) was devised for writing Mandarin. (In 1937 it received formal backing from the government, but World War II stopped further progress.) In 1929 a National Romanization, worked out by the author and language scholar Lin Yü-t'ang, the linguist Yuen Ren Chao, and others, was adopted. This attempt also was halted by war and revolution. A rival Communist effort known as *Latinxua*, or Latinization of 1930, fared no better. An attempt to simplify the language by reducing the number of characters to about 1,000 failed because it did not solve the problems of creating a corresponding "basic Chinese" that could profitably be written by the reduced number of symbols.

The government of China has taken several important steps toward solving the problems of the Chinese writing system. The first and basic step of making one language, Modern Standard Chinese, known throughout the country has been described above. In 1956 a simplification of the characters was introduced that made them easier to learn and faster to write. Most of the abridged characters were well-known unofficial variants, used in handwriting but previously not in printing; some were innovations. In 1958 the previously mentioned romanization known as *pinyin zimu* was introduced. This system is widely taught in the schools and is used for many transcription purposes and for teaching Modern Standard Chinese to non-Chinese peoples in China and to foreigners. Pinyin in theory is conceived as a script that will gradually replace characters. (For information on Chinese as a writing system, see WRITING: *Chinese writing*. For information on Chinese calligraphy, see WRITING: *Chinese calligraphy*.)

Reconstruction of Chinese protolanguages. For reconstructing the pronunciation of older stages of Sinitic, the Chinese writing system offers much less help than the alphabetic systems of such languages as Latin, Greek, and Sanskrit within Indo-European or Tibetan and Burmese within Sino-Tibetan. Therefore, the starting point must be a comparison of the modern Sinitic languages, with the view of recovering for each major language group the original common form, such as Proto-Mandarin for the Northern languages and Proto-Wu and others for the languages south of the Yangtze River. Because data are still lacking from a great many places, the once-standard approach was to compare major representatives of each group for the purpose of reconstructing the language of the important dictionary *Ch'ieh-yün* of AD 601 (Sui dynasty), which mainly represents a Southern language type. One difficulty is that the language in a given area represents a mixture of at least two layers: an older one of the original local type, antedating the language of the *Ch'ieh-yün*, and a younger one that is descended from the *Ch'ieh-yün* language or a slightly younger but closely related tongue—the so-called T'ang koine, the standard spoken language of the T'ang dynasty. The relationship of the protolanguages is further complicated by the different substrata of non-Chinese stock that underlie many if not most of the major languages.

Mixture of language layers

The degree to which the Sinitic languages have been influenced by the T'ang (or Middle Chinese) layer varies. In the North the Old Chinese layer still dominates in phonology; in Min the two layers are kept clearly apart from each other, and the Middle Chinese layer is most important in the reading pronunciation of the characters; Yüeh has two Chinese layers of the Southern type and is typologically similar to a Tai substratum.

The Old Chinese layer is characterized by early decay of final consonants, late development of tones from sounds or suprasegmental features located toward the end of the syllable, change of final articulation type because of similar initial type (as in syllables with more than one voiced activity, which may change or lose one of these; phenomena later manifested as a tonal change), and influence of sounds and tones in a syllable on those of surrounding ones (sandhi).

The New Southern stratum in Sinitic languages is characterized by early change of final articulation types into tones, extensive development of registers according to type of initial consonant, and late or no loss of final stops. The Old layer cannot be the direct ancestor of the New layer. The division into Northern and Southern dialects must be very old. It might be better to speak of a T'ang and a pre-T'ang layer, or a T'ang and a Han layer (the Han dynasty was characterized by extensive settlement in most parts of what is now China proper).

The Ch'ieh-yün dictionary. For a long time the *Ch'ieh-yün* dictionary was assumed to represent the language of the capital of the Sui dynasty, Ch'ang-an (in the present province of Shensi), but research has demonstrated that its major component was the language of the present-day Nanking area with a certain attempt at compromise with other speech habits. As its first criterion for classifying syllables, the *Ch'ieh-yün* dictionary takes the tones, of which it has four: *p'ing, shang* (here transcribed with a colon, as in *pa:*), *ch'ü* (here transcribed with a hyphen, as in *pa-*), and *ju*, or even, rising, falling, and entering ("checked") tones. The entering tone comprised those syllables that ended in a stop (*-p, -t, -k*). The rising and falling tones may have retained traces of the phonetic conditioning factor of their origin, voiced and voiceless glottal or laryngeal features, respectively. The even tone probably was negatively defined as possessing no final stop and no tonal contour.

Tones of the *Ch'ieh-yün* dictionary

Next, the dictionary is divided according to rhymes, of which there are 61, and, finally, according to initial consonants. Inside each rhyme an interlocking spelling system known as *fan-ch'ieh* was used to subdivide the rhymes. There were 32 initial consonants and 136 finals. The number of vowels is not certain, perhaps six plus *i* and *u*, which served also as medial semivowels. The dictionary contained probably more vowels than either Archaic Chinese or Modern Standard Chinese, another indication that the development of the Northern Chinese phonology did not pass the stage represented by *Ch'ieh-yün*.

Additional sources. There are additional sources for reconstructing the *Ch'ieh-yün* language: Chinese loanwords in Vietnamese, Korean, and Japanese (Japan has two different traditions—Go-on, slightly older than *Ch'ieh-yün* but representing a Southern language type like *Ch'ieh-yün*, and Kan-on, contemporary with *Ch'ieh-yün* but more similar to the Northern tradition) and Chinese renderings of Indo-Aryan (Indic) words. Voiced stops are recovered through Wu, Hsiang, and Go-on (*e.g.*, Modern Standard Chinese *t'ien* "field," Wu and Hsiang *di*, Go-on *den*, *Ch'ieh-yün dhien*), final stops especially through Yüeh and Japanese (*e.g.*, Modern Standard Chinese *mu* "wood," Yüeh *muk*, Go-on *mok*[*u*], *Ch'ieh-yün muk*), and retroflex initial sounds from Northern Chinese (*e.g.*, Modern Standard Chinese *sheng* "to live," *Ch'ieh-yün ṣʌŋ* [the *ṣ* is a retroflex]).

Early Archaic Chinese is the old stage for which the most information is known about the pronunciation of characters. The very system of borrowing characters to write phonetically related words gives important clues, and the rhymes and alliteration of the *Shih ching* furnish a wealth of details. Even though scholars cannot always be sure that prefixes and infixes are correctly recovered, and though the order in which recoverable features were pronounced in the syllable is not always certain (*rk-* or *kr-*, *-wk* or *-kw*, and so on), enough details can be obtained to determine the typology of Old Chinese and to undertake comparative work with the Tibeto-Burman and Karenic languages. The method employed in this part of the reconstruction of Chinese has been predominantly internal reconstruction, the use of variation of word forms within a language to construct an older form. As knowledge of the old layer of modern languages and dialects increases, however, the comparative method, which draws on similarities in several related tongues, gains importance. Through further internal reconstruction, features of the Proto-Sinitic stage, antedating Archaic Chinese, can then be restored.

TIBETO-BURMAN LANGUAGES

The Tibeto-Burman languages have evolved from the ancestral language, Proto-Tibeto-Burman, in vastly different ways and at their own pace, in accordance with the geographic and social factors that have determined the fate of Central and South Asian peoples. Some tribes have been stationary; others have swept over huge areas. As a result, conservative or archaic features do not occur in only one contiguous part of the language area and innovations in another. The nearest genetic relations are often not identical with the closest typological ones.

Tibetan. Of the modern Tibetan languages and dialects, the Western ones have preserved initial consonant clusters and final stops most faithfully and have had the least compensational development of tones. Most Central languages and dialects, including Lhasa, have lost all consonant clusters and final stops and in the process have acquired a larger inventory of single consonants and a system of tones. These changes and reductions are linked to a similar reshaping of certain grammatical processes of word formation that now operate only through suprasegmental and syllabic elements. To a surprising degree, however, Modern Central Tibetan possesses grammatical categories identical with or very similar in content, though not in form, to those of Classical Tibetan (Modern Standard Chinese bears a similar relationship to Old Chinese). The relationship of nouns to the main verb is indicated through postposed particles, the agent of a transitive verb indicated as the one by whom the action is performed, and the subject of an intransitive verb expressed in the same way as the object or goal of the action (so-called ergative constructions). Nominal modifiers precede nouns, and verbal modifiers follow verbs. The main verb, always placed after all nouns, is followed by particles expressing aspect and tense.

Old Tibetan pronunciation can be reconstructed by comparison of modern dialects and through the very conservative alphabetic script of Indian origin that goes back to the 7th century AD and found its present form in the 9th century. The orthography is far removed from present-day Standard Tibetan pronunciation.

Old Tibetan is one of the most archaic of the Tibeto-Burman languages. It retained Tibeto-Burman final stops and final *-r*, *-l*, *-s* and also the initial voiced consonants. Many Old Tibetan consonant clusters may be referred to as Proto-Tibeto-Burman. The case particles and complicated verbal conjugation perhaps represent an elaboration on somewhat simpler tendencies in the protolanguage.

Himalayish languages

Some Himalayish languages are in certain respects as archaic as Tibetan, although most initial clusters are gone. An old feature is the connection of voiced–voiceless initial consonants with intransitive–transitive verbs. Because they have developed the feature of incorporating agent and object pronouns in verbs (and of possessive pronouns in nouns), these have been known as "pronominalizing" languages. An influence from contiguous Indo-European languages seems possible, but not certain.

Some Kirantish languages have retained consonant clusters and voiced initial consonants; others have given up both. Bahing has maintained the connection between voicedness and transitivity. Within Mirish, which has kept voiced initial sounds, Abor retains final stops and Dafla has some initial consonant clusters. Kachinish is quite conservative; prefixes are well retained as are voiced initial consonants, although some reshuffling has taken place in this respect.

Burmese. Within Burmish, Modern Standard Burmese has undergone a set of radical changes. Initial *ts-* and *tsh-* have become *s-* and *sh-*; *s-* has become *θ-* (pronounced *th* as in "thin"); and *y-* and *r-* have coalesced as *y-*, and *ky-* and *kr-* as a palatal *c* (*ty*). Furthermore, all final consonants except nasals have coalesced as glottal stops, and all nasals have resulted in nasalization of the preceding vowel. Also, the quality of vowels has been greatly altered. As was the case in Tibetan, in spite of great phonetic changes, grammatical categories are similar to those that scholars envisage for Proto-Tibeto-Burman. Cases of nouns and aspects of verbs are expressed through postposed particles.

Among the Burmese dialects, Arakanese is especially conservative, and the closely related language Maru is one of the most archaic within Tibeto-Burman in respect to final consonants. The Yi languages lost most consonant clusters, as did all Burmish languages, but tend otherwise to be conservative in their treatment of initial sounds and radical in the loss and change of final consonants. Nung has retained final stops lost in Burmish (*-r*, *-l*, *-s*) and possesses a set of prefixes like those of Kachin.

The Burmese writing system

Study of the conservative Burmese writing system, in combination with comparative linguistic work, makes possible the reconstruction of Old Burmese. The language of the Myazedi inscription of 1113 is similar in its sound system to written Burmese in its present form, which dates to at least the 15th century. The writing system was taken over from the Mon people, who had developed their writing from Pyu, a Sino-Tibetan language known in Myanmar from approximately AD 500. It is alphabetic of an Indian type but represents a separate Southern line of development.

Old Burmese is phonetically further from Proto-Tibeto-Burman than is Tibetan. Initial clusters have mostly disappeared but are felt in the development of initial consonants. Some clusters with *-w-* and liquid sounds have been retained. The tonal system of Burmese (unlike that of Tibetan) developed to compensate for the loss of final features.

The Baric languages maintain a few older prefixes and developed some of their own. They tend to retain or merge *r* and *l*. The relationship of voiced–voiceless and intransitive–transitive is retained, but a significant reshuffling of initial consonants took place, as it did also in Meithei and Kukish. The Kukish languages are found in all stages of development, but many of them are among the most archaic and most important for the reconstruction of Tibeto-Burman. Prefixes are best preserved in Old Kuki and in the Nāgā branch, whereas vowels seem very archaic in Lushai of the Central branch (the Lushai vowels have a difference of length that must in some way be explained in terms of Proto-Tibeto-Burman). Some Kukish languages incorporate pronouns in verbs. (S.C.E.)

TAI LANGUAGES

The name Tai denotes a family of closely related languages, of which the Thai (Siamese) language of Thailand is the most important member. Because the word Thai has been designated as the official name of the language of Thailand, it would be confusing to use it for the various other languages of the family as well. Tai is therefore used to refer to the entire group.

THE DISTRIBUTION AND CLASSIFICATION OF TAI LANGUAGES

Geographic distribution

Spoken in Thailand, Laos, Myanmar (Burma), Assam in northeastern India, northern Vietnam, and the southwestern part of China, the Tai languages together form an important group of languages in Southeast Asia. In some countries they are known by different tribal names or by designations used by other peoples. For example, there is Shan in Myanmar; Tai in Yunnan, China (includes languages known outside China as Nüa and Lü); Chuang in Kwangsi, China; Pu-i in Kweichow, China; Tay, Nung, White Tai, Black Tai, Red Tai, and others in northern Vietnam; and Khün, Lü, and others in Thailand and Laos. Designations vary over time as well: older names include Pai-i (for Tai); Chuang-chia (Chuang); Chung-chia, Dioi, Jui, and Yai (Pu-i); and Tho, which is still sometimes used for the language or languages now known in Vietnam as Tay. Ahom, an extinct language once spoken in Assam (India), has a considerable amount of literature. The Tai languages are divided into three linguistic groups—the Southwestern, the Central, and the Northern. Thai and Lao, the official languages of Thailand and Laos, respectively, are the best known of the languages.

The number of Tai speakers is estimated at 80 million. Of these, about 55 million are in Thailand, some 18 million in China, and about 7 million in Laos, northern Vietnam, and Myanmar. There are tremendous variations between several estimates, and these figures may serve as only rough indications of the Tai populations.

The relationship of Tai languages to other language families. The closest relatives of the Tai languages are the Kam-Sui languages, spoken in Kweichow, China; and the languages Be (Hainan Island) and Lakkia (Kwangsi, China). More distantly related are a number of languages and language groups sometimes given the collective name Kadai. They include the Li, or Hlai, languages of Hainan; the Ch'i-lao (Kelao, or Gelao) languages spoken mostly in Kweichow; Laha of northern Vietnam; Lachi (formerly Lati); Laqua (also called Ka-beo and Pubiao, Pinyin Qabiau); and others. Hlai has just over a million speakers, but the other languages have much smaller populations, and all are endangered to some degree. The entire language family containing Tai and all its relatives is called either Tai-Kadai or simply Kadai.

It was formerly assumed that Tai and its relatives belonged to the Sino-Tibetan family because of similar phonological systems (especially tone) and many lexical items shared with Chinese. But many more lexical items are unrelated to Chinese, and, since these include the preponderance of the most basic vocabulary, a Sino-Tibetan affiliation is not tenable. A competing proposal links Tai and its relatives with Austronesian, but this connection has not yet been established to the satisfaction of most scholars.

Classification within the family. *Criteria for classification.* The classification used for this article is based on linguistic relationships proposed in 1959–60; the criteria for it are lexical (involving similarities in vocabulary) and phonological (involving similarities in sounds and systems of sounds). According to these features the Tai languages are divided into the three groups mentioned above (see map). Languages of the Southwestern group are spoken in Thailand, Laos, northern Vietnam, Myanmar, and Yunnan, China; they include Thai, Lao, Shan, Khün, Lü, White Tai, Black Tai, and others. The Southwestern division, which is geographically the most widespread group, consists of two-thirds of the Tai-speaking population and represents an expansion that has occurred in comparatively recent periods. To the Central group belong the Tho dialects that are spoken in northern Vietnam and the various dialects spoken in Kwangsi, such as Lungchow. The Pu-i dialects in Kweichow and the Chuang dialects in Kwangsi belong to the Northern group.

Main groups of Tai

Differences in vocabulary. The fairly large number of vocabulary items that are shared by languages of these three groups suggests their genetic relationship. In some instances, however, there are items that are shared by only two of the groups and are not found in the other. For instance, the word for 'sky' is shared by the Southwestern dialects (Thai *fáa*) and the Central dialects (Lungchow *faa*), but another word is used in the Northern dialects (Po-ai *mɯn*). Similarly, the word for 'beard' is shared by the Central group (Lungchow *mum*) and the Northern group (Po-ài *mum*) but is replaced by another word in the Southwestern group (Thai *nùat*). In another instance the term for 'knife' is shared by the Southwestern (Thai *mîit*) and the Northern groups (Po-ai *mit*) but not by the Central dialects, which have various words such as Lei Ping *taau* and Ning Ming *pjaa* (Lungchow has both *pja* and *taau*). There are also vocabulary items that are found only in one of the three groups. The evidence seems to indicate that there are three groups of dialects in the Tai family.

Differences in phonology. Different phonological features may be reconstructed for the ancestral form of some words according to the dialect group. For instance, the Southwestern forms for the verb 'to be' (Thai *pen*) are derived from a protoform **pɛn* (with the vowel pronounced as in English *egg*), whereas the Central dialect forms (Lungchow *pin*) and the Northern forms (Po-ai *pan*) come from a protoform **bɛn*. (A protoform is the presumed or reconstructed ancestral form of a word; an asterisk [*] indicates an unattested, reconstructed form.) Similarly, the Southwestern and the Central forms for the classifier for animals (Thai *tua,* Lungchow *tuu*) are derived from a protoform **tua,* whereas the Northern forms (Po-ai *tuu*) are attributed to a protoform **dua.* (A classifier is a term that indicates the group to which a noun belongs [for example, 'animate object'] or designates countable objects or measurable quantities, such as 'yards [of cloth]' and 'head [of cattle]'.) Such words as the forms for 'to be' and the classifier for animals are good indications of dialect boundaries.

Dialect groups

In phonological development, the Northern dialects differ from the rest in not maintaining the distinction between aspirated and unaspirated voiceless stops. That is, the dialects have lost the feature of aspiration, which sounds like a puff of breath accompanying a consonant. Aspiration may, however, be reintroduced in some dialects by later borrowing or secondary developments. The Central dialects differ from the other groups in the treatment of certain Proto-Tai consonant clusters, such as **tr-* and **thr-*. Although they have changed from the protoforms, these are usually kept distinct in the other groups—*e.g.,* in Thai as *taa* ('eye') and *haaŋ* ('tail'), in Po-ai as *taa* and *lɯɯŋ*. In the Central dialects, however, they have merged into a single sound—*e.g.,* Tho *thaa* and *thaaŋ,* Lungchow *haa* and *haaŋ.*

LINGUISTIC CHARACTERISTICS

Phonological characteristics. The sound system of the Tai languages may best be described in terms of its syllabic structure. Each syllable consists of an initial consonant or consonant cluster followed by a vowel or vowel cluster (long vowel or diphthong), which may be further followed by a final consonant, usually a nasal sound or an unreleased stop. (An unreleased stop is a consonant in which there is complete stoppage of the airstream from the lungs and in which the tongue or the lips maintain the position of the consonant without opening the stoppage.) In addition, each syllable has a tone. As an illustration of this type of structure, Table 46 shows the system of Thai, a national standard based on the language spoken in the Bangkok area.

Table 46: The Sounds of the Thai Language

	initial consonants*					final consonants	
Labials	p	ph	b	m	f	p	m
Alveolars	t	th	d	n		t	n
Sibilants	c	ch			s		
Velars	k	kh		ŋ		k	ŋ
Glottals	ʔ				h	ʔ	
Semivowels		j	w			j	w
Liquids		l	r				

	vowels**			long vowels			diphthongs		
High	i	ɯ	u	ii	ɯɯ	uu	ia	ɯa	ua
Mid	e	ɤ	o	ee	ɤɤ	oo			
Low	ɛ	a	ɔ	ɛɛ	aa	ɔɔ			

*Certain consonant clusters are also permitted initially, such as *kl, kr, khl, khr, tr, pl, pr, phl, phr, kw,* and *khw.* **The vowel symbols /a e i o u/ are pronounced as in Italian or Spanish. Others are ɛ (as in *bat*), ɔ (as in *law*), ɯ (*u* pronounced with lips spread rather than rounded), and ɤ (*o* pronounced with lips spread rather than rounded).

Thai has five tones: level (using no diacritic), low (using a grave accent [`]), falling (using a circumflex [^]), high (using an acute accent [´]), and rising (using a wedge, or haček, [ˇ]); for example, *maa* 'to come,' *màak* 'areca nut,' *mâak* 'much,' *máa* 'horse,' and *mǎa* 'dog' are differentiated by various tones. Five tones

Grammatical characteristics. The statements about morphology and syntax refer particularly to the standard language of Thailand, though they are applicable in general to all the Tai languages.

Morphology. Words or morphemes (word elements) are, for the most part, monosyllabic, but there are also many polysyllabic words, mainly compounds and loanwords from the Indic languages Sanskrit and Pāli and from Khmer, an Austroasiatic language spoken mainly in Cambodia. There are no inflections in Tai comparable to *-ed* in English 'walked' or *-s* in 'dogs.' The chief Tai method of forming new words is compounding—*e.g., nâa-taa* 'countenance' (literally, 'face-eye'), *kèp-kìaw* 'to harvest' (literally, 'gather-cut with a sickle'). Reduplication, the repetition of a word or part of it, is common—*e.g., dîi-dii* 'very good' from *dii* 'good.' Partial reduplication is also found, such as *sanùk-sanǎan* 'to enjoy oneself' from *sanùk* 'to have fun.' Prefixes and infixes occur often in Indic and Khmer loans—*e.g., pra-thêet* 'country' (from Sanskrit *pradeśa*) and *d-amn-ɤɤn* 'to proceed,' derived from *dɤɤn* 'walk' (from Khmer *damnaɤ* 'gait, bearing' and *dae* 'walk,' respectively). There also developed some native prefixes that are abbreviated forms of what once were full words. For example, *ma-,* a prefix used in many names of fruits, such as *ma-phráaw* 'coconut,' *ma-mûaŋ* 'mango,' is derived from *màak* 'areca nut' (originally 'fruit'). Similarly, in the word *sa-dɯɯ* 'navel,' *sa-* is the reduced form of *sǎaj* 'line, string,' which refers to the umbilical cord. Old processes of derivation involved using the alternation of consonants or tone or both, such as *níi* 'this' and *nîi* 'here,' *nɔ̌ɔj* 'small, little' and *nɔ̀ɔj* 'a little bit,' *khiaw* 'sickle' and *kìaw* 'to cut with a sickle.' Such processes are, however, no longer active.

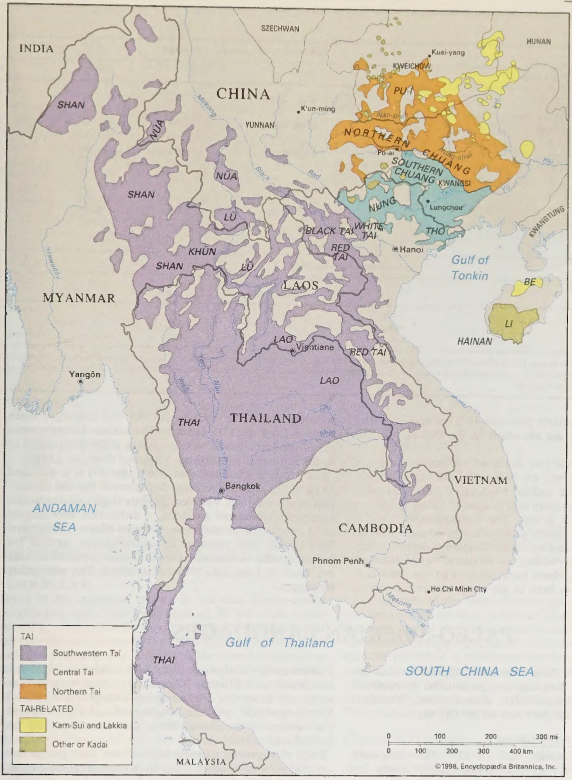


Figure 25: Major divisions of the Tai languages and related languages.

Syntax. Word order is generally subject–verb–object (SVO). The verb plus object may be called the verb phrase, which functions as predicate. Nouns also may be used as predicates; an example is *wan-níi wan-sǎw* 'today (is) Friday.' A noun phrase consists of a noun, which may be followed by its modifiers (another noun, an adjective, or a verb phrase), which, in turn, are followed by a numeral with a classifier and, finally, by a demonstrative. For example, *tûu* 'cabinet' and *nǎŋsɯ̌ɯ* 'book' become *tûu-nǎŋsɯ̌ɯ* 'bookcase'; *tûu-kèp nǎŋsɯ̌ɯ* is also 'bookcase' ('cabinet-keep-book'); *tûu-yen* is 'icebox' ('cabinet-cool'); and *tûu-kèp-nǎŋsɯ̌ɯ sɔ̌ɔŋ-baj níi* is 'these two bookcases' ('cabinet-keep-book two-classifier these'). Classifiers must be used with numerals and vary according to the nouns they are to enumerate. It is not uncommon for a noun to serve as its own classifier.

A verb phrase may consist of an adjective or of a verb, often followed by its object or its complements or both. A common type of complement denotes the direction of action by using such words as *paj* 'to go,' *maa* 'to come,' *khɯ̂n* 'to ascend,' *loŋ* 'to descend,' and so on. Examples are *ʔaw nǎŋsɯ̌ɯ paj* 'Take the book away!'; *ʔaw nǎŋsɯ̌ɯ maa* 'Bring the book (here)!'; *aw nǎŋsɯ̌ɯ khɯ̂n maa* 'Bring the book up!'; and *aw nǎŋsɯ̌ɯ loŋ paj* 'Take the book down!'. A number of particles, particularly those occurring at the end of a sentence, are used to indicate a question or a command, to express emphasis or uncertainty, to show politeness and the sex of the speaker, and so on. For example, *khâw maa lɛ́ɛw* 'He has come'; *khâw maa máj* 'Is he coming?'; *khun maa máj khráp* 'Are you coming?' (man speaking, polite); *khun maa máj khá* 'Are you coming?' (woman speaking, polite).

Vocabulary. The various Tai languages have been in contact with many different languages spoken in the same area of Southeast Asia, and it is inevitable that words from different sources have been adopted by them. There is a basic core of vocabulary that is shared by most of the Tai languages and may be considered as the native vocabulary of the protolanguage. Nevertheless, items in this list may be found to resemble forms in other languages, such as Chinese or Indonesian, and they have given rise to different interpretations of the relationship of the Tai languages to other linguistic families. Loanwords in later Loanwords
periods from the Khmer and Indic languages (Sanskrit

Table 47: The Thai Alphabet

consonants

middle class		high class			low class				
	ก k	ข kh			ค kh		ฆ kh	ง ŋ	
	จ c	ฉ ch			ช ch	ซ s	ฌ ch	ญ j	
ฎ d	ฏ t	ฐ th			ฑ th		ฒ th	ณ n	
ด d	ต t	ถ th			ท th		ธ th	น n	
บ b	ป p	ผ ph	ฝ f		พ ph	ฟ f	ภ ph	ม m	
		ศ s	ษ s	ส s	ย j		ร r	ล l	ว w
	อ ʔ	ห h			ฮ h		ฬ l		

diphthongs, vowel-consonant combinations		vowels*	
short	long	short	long
◌ัวะ -uaʔ†	◌ัว -ua	-ะ -aʔ	-า -aa
	-ว- -ua-	◌ั -a-	
เ◌ียะ -iaʔ†	เ◌ีย -iaʔ	◌ิ -iʔ	◌ี -ii
	เ◌ือ -ɯa	◌ึ ɯʔ	◌ือ -ɯɯ
ไ, ใ -aj	-าย -aaj	◌ุ uʔ	◌ู -ɯɯ
เ-า -aw	-าว -aaw	เ-ะ -eʔ	เ- -ee
	เ-ย -ɤɤj	แ-ะ -ɛʔ	แ- -ɛɛ
◌ำ -am		โ-ะ oʔ	โ- -oo
ฤ rɯ(ʔ)	ฤ rɯɯ	◌ -o-	
ฦ lɯ(ʔ)	ฦ lɯɯ	เ-าะ ɔʔ-	อ -ɔɔ
		เ-อะ -ɤʔ	เ-อ -ɤɤ
			อ— -ɤɤ-

tone mark and syllable type

consonant class	no mark	◌่	◌้	checked long	checked short	◌๊	◌๋
high	rising	low	falling	low	low	(not used)	
middle	mid	low	falling	low	low	high	rising
low	mid	falling	high	falling	high	(not used)	

*Pronunciation key: ɔ=*aw* as in *law;* ɤ=*a* as in *sofa;* j=*y* as in *yawn;* ʔ=glottal stop (the catch sound for *tt* in Cockney *bottle*, or the consonant separating the syllables of *uh-oh*). †Indicates rare or obsolescent. ††Checked syllables are those having a final *p, t,* or *k* or those having a short vowel and no final consonant. Tone marks are generally not used with checked syllables, but with other syllable types (with final long vowel or diphthong, or with finals *m, n, ŋ; j, w*).

and Pāli) are particularly common in Thai and Lao, and loans from Chinese are abundant in the Central and the Northern dialects.

Writing. There are two kinds of writing used among the Tai languages. One, ultimately derived from Chinese, is used by the Central and the Northern dialects; the other comes from Indic sources and is used in many languages of the Southwestern group. The Chinese-based system, which is employed chiefly to write songs, consists of both Chinese characters and modified Chinese characters, very much like the early writing in Vietnam. A specimen of this type of writing dates back perhaps to the 18th century, but it may have been in use much earlier. The script adapted from Indic sources for the Tai languages dates perhaps from the 13th century. The earliest known example of the Indic-based writing system is the inscription of Ramkhamhaeng in northern Thailand from AD 1293.

The Modern Thai alphabet (see Table 47) is a modified form of the original writing. It preserves the old distinction of voiced (low), voiceless aspirate (high), and voiceless unaspirate/glottalized (middle), a distinction now largely lost but one that nevertheless leaves its effects on the tone. This system also provides an unambiguous method for indicating the vowels and tones. Similar types of writing are used in Lao, Lü, White Tai, Black Tai, and other languages. (F.K.L./D.B.So.)

PALEO-SIBERIAN LANGUAGES

The collective term Paleo-Siberian is applied to four genetically unrelated language groups situated in northern Asia—Yeniseian, Luorawetlan (Luoravetlan), Yukaghir (Yukagir), and Nivkh (formerly called Gilyak).

THE LANGUAGES OF THE GROUP

Yeniseian, Luorawetlan, and Nivkh. The Yeniseian group is spoken in the Turukhansk region along the Yenisey River. Its only living members are Ket (formerly called Yenisey-Ostyak), which is spoken by about 500 persons, and Yug, with no more than 5 speakers. Kott (Kot; also called Assan or Asan), Arin, and Pumpokol, now-extinct members of this group, were spoken chiefly to the south of the present-day locus of Ket and Yug.

The Luorawetlan family consists of (1) Chukchi, spoken by no more than 11,000 people in the northeasternmost parts of Siberia, west of the small enclaves of Siberian Yupik (Eskimo), (2) Koryak, also called Nymylan, with approximately 3,500 speakers, spoken on northern Kamchatka and northward to the Anadyr River basin, (3) the strongly divergent but probably related Itelmen (or Kamchadal), with a bare remnant of 500 speakers on the central west coast of Kamchatka, (4) Aliutor, perhaps a Koryak dialect, with about 2,000 speakers, and (5) Kerek, with about 10 speakers.

Yukaghir. Yukaghir (regional name Odul) is spoken by about 200 persons (less than 20 percent of the ethnic group) who are divided about equally into two enclaves: Tundra Yukaghir (also called Northern Yukaghir) in the Sakha republic (Yakutia), near the estuary of the Indigirka River; and Kolyma, or Forest, Yukaghir (also called Southern Yukaghir) along the bend of the Kolyma River. Extinct earlier dialects or languages related to Yukaghir are Omok and Chuvan (Chuvantsy); these were spoken south and southwest of the current Yukaghir area. Nivkh has about 1,000 speakers, roughly half of whom live in the estuary of the Amur River and the other half on Sakhalin Island.

Paleo-Siberian groups not genetically related

Lack of a genetic relationship. These four groups are not related to each other. They have been subsumed under the names Paleo-Siberian, Paleo-Asiatic, or, more rarely, Hyperborean, ever since the Baltic German zoologist and explorer Leopold von Schrenck surmised, in the middle of the 19th century, that they constituted the remnants of a formerly more widely dispersed language family that had been encroached upon by invading groups of Uralic and Altaic speakers. Schrenck's hypothesis is quite correct to the extent that as recently as the 17th century Yeniseian, Luorawetlan, and Yukaghir languages were spoken over much wider territories than they are today. For example, it is known that Samoyed languages (of the Uralic family) at one time in the past absorbed the languages of now-extinct Yeniseian tribes, that Yukaghir was spoken as far west as the Taymyr Peninsula in the 17th century, and that the former domains of Chukchi and Koryak extended much farther to the west. Little is known about the prehistory of Nivkh, but it may be assumed that this language was also originally centred farther to the west, perhaps in Manchuria. As far as can be determined with the help of the methods of comparative linguistics, however, the four present-day Paleo-Siberian groups never formed a single family of languages in the accepted sense of that term. In fact, they may represent only a fragment of a possibly greater diversity of language families in prehistoric Siberia.

Figure 26: Distribution of Paleo-Siberian languages.

Many of the languages spoken in the area during earlier periods may have been swallowed up by the more recent as well as more culturally vigorous intruders in Siberia that are now the neighbours of the Paleo-Siberian enclaves; this includes primarily the Yakut (whose domains stretch as far as the Chukchi and Yukaghir areas) and also various Tungus tribes (one or another of which borders on each of the Paleo-Siberian languages).

The Paleo-Siberian languages and other language families. Many attempts have been made to show that the four Paleo-Siberian families are related either to each other or to adjacent (or more distant) language families. Thus, Ket has been compared with the Sino-Tibetan family (which includes Chinese and Tibeto-Burman languages) and with some of the languages of the Caucasus, and Yukaghir has been compared with Uralic. Some of these comparisons (*e.g.*, the comparisons of Ket with Caucasian languages) are fanciful experiments or completely unfounded. Even more sober efforts to demonstrate ties with other languages are seriously hampered by many millennia of separation from possible related languages. The systematic reconstruction of protolanguages, often assumed as a standard for relationship among the better-established language families, is not available. Of attempts to find relatives for the four Paleo-Siberian families, only a Uralo-Yukaghir relationship has been positively received. The remaining Paleo-Siberian families, including Luorawetlan as a family in its own right, must continue to be regarded as isolates, unrelated to any known language. Although numerous resemblances in grammatical or phonological traits may be observed between Paleo-Siberian and adjacent languages (such as between Chukchi and Yupik, between Samoyed and Yukaghir, or between Nivkh and Korean or Japanese), these are not indexes of genetic affinity but are often the result of the diffusion of linguistic traits over large geographic areas. They may, however, provide clues to the linguistic prehistory of Siberia.

The cultures of the Paleo-Siberian groups are similar in that they are all Arctic or subarctic. Each particular group, however, has its own characteristic cultural profile. These characteristics may on occasion very closely resemble the cultural profile of a non-Paleo-Siberian group; *e.g.*, Ket culture resembles Selkup culture (the Selkup language is classified as Uralic) more closely than it resembles that of any Paleo-Siberian group, evidently because Selkup- and Ket-speaking groups are located in contiguous areas.

LINGUISTIC FEATURES

Phonological features. Typical phonological features of the Paleo-Siberian languages are postvelar consonants (*i.e.*, sounds that are formed farther back in the mouth than /k/ and usually represented as *q*), vowel harmony of various kinds (*e.g.*, the alternation of *e* and *i* in the form for 'my' in Nivkh *ñe-řla* 'my harpoon' and *ñi-řly* 'my sky'), consonant alternations (*e.g.*, the alternation between *b*, *v*, and *f* in Nivkh *bal* 'mountain,' *ñ-val* 'my mountain,' *c-fal* 'your mountain'), and rich consonant clusters in all but Yukaghir, but especially in Nivkh (*e.g.*, South Sakhalin Nivkh *tfyfř* 'hammer') and Itelmen (*e.g.*, *ktxam* 'bone').

Grammatical features. The grammatical structures of the Paleo-Siberian groups differ considerably from each other. In a broad sense, Nivkh resembles Japanese in its grammatical categories and processes (in word order, heavy inflection of verbs, and use of enclitics—an enclitic is a word that is associated with a preceding word), whereas Yukaghir shares certain grammatical categories with some Uralic languages—for instance, the use of distinct conjugations to focus on the roles of major sentence categories (*e.g.*, subject, object, negation): Yukaghir *tet mer-ai-mek* 'you *shot*' versus *tet el-ai-yek* 'you did *not* shoot' versus *tet ai* '*you* shot' versus *tet ileleŋ ai-me-ŋ* 'you shot *at the deer*.' A typical feature of Luorawetlan is its strong tendency toward complex compounding (also called incorporation) and circumfixation; for example, in

Chukchi *ga* + *mor-ïk* + *tor* + *orw-ïma* 'in our new sleigh,' the entire unit is surrounded by the circumfix *ga-* . . . *-ïma* 'in' (compare *ga* + *mor-ïk* + *orw-ïma* 'in our sleigh,' without *tor* 'new,' and *ga* + *tor* + *orw-ïma* 'in the new sleigh,' without *mor-ïk* 'our'). A characteristic feature of the Ket verb is its succinct complexity, involving such categories as gender, animateness, and type of event; for instance, *t-k-it-n-a* 'I carved it up,' which consists of *t-* 'I,' the verbal complex *-k-* . . . *-a* 'cut up (carve, split) into pieces once,' *-it-* (feminine object marker 'her, it'), and *-n-* (past-completed tense).

All the Paleo-Siberian languages are quite rich in devices that can be used for compounding words. In syntax, Luorawetlan favours ergative constructions in which markers indicate the agent or instrument of the action—*e.g., Father* + agent marker, *bear* (subject), *shoot* (main verb), 'Father is shooting a bear.'

Vocabulary. In addition to the stock of native words inherited from its ancestral language, each Paleo-Siberian language also has numerous loanwords, some of which are recent and from adjacent or recently adjacent languages and others of which are ancient and from languages with which it no longer has contact. Some of the loanwords from ancient times are consequently more difficult to identify and trace to their origins. In general, Manchu-Tungus (or Tungus), a subfamily of the Altaic language family, is the source of most loanwords, but the Turkic languages (including Yakut) have also served as the sources of loans, and Ket has borrowed some words from Selkup. There are also more complicated loan relationships, such as those that are found in the reindeer terminology of Nivkh, which is borrowed from a Manchu-Tungus language but seemingly not from any of those with which Nivkh is now in contact. South Sakhalin Nivkh also contains a considerable number of loanwords from Ainu (a language of northern Japan) and, during the first half of the 20th century, was hospitable to potential loans from Japanese; the Japanese loanwords never became acculturated because the Japanese hegemony over South Sakhalin ceased after 1945. Chukchi has some Yupik loanwords.

Loanwords in Paleo-Siberian languages

The most viable source of technical and of all the other modern vocabulary has been the Russian language, the influence of which began with the first contact and continues to be strong. Each Paleo-Siberian language adjusts the Russian loanwords according to the dictates of its phonology and grammar, but the more recently borrowed words tend to retain their original Russian form or one closely resembling it.

Writing. The Yukaghir had not a system of true writing, but rather a pictographic tradition (incisions on fresh birch bark) used by men for route maps and by young women as an aid in the public expression of feelings on the themes of love and separation. Limited use of such a system among Koryak speakers has also been reported.

Since the 1920s and '30s each Paleo-Siberian language has had a literary language and a script now based on the Cyrillic alphabet (and formerly based on the Latin script). Because at one time these native languages were used in part in elementary education, primers and arithmetic books for the lowest grades were available. Some natives continue their education and acquire a good knowledge of Russian culture. This situation has led to the rise of bilingualism, but it also has contributed to the growth of modern literatures in the native languages, based on Russian models, especially among the Koryak and the Chukchi.

(R.Au./D.M.Ab.)

CAUCASIAN LANGUAGES

The term Caucasian languages as used in this article includes groups of languages indigenous to Transcaucasia and adjacent areas of the Caucasus region, between the Black and Caspian seas; it excludes the Indo-European (Armenian, Ossetic, Talysh, Kurdish, Tat) and Turkic languages (Azerbaijani, Kumyk, Noghay, Karachay, Balkar) and some other languages of the area, all of which were introduced to the Caucasus in historical times. The Caucasian languages are also referred to as Paleo-Caucasian and Ibero-Caucasian languages.

The Caucasian languages are found in the territory north and south of the Great Caucasus range; their number varies, according to different classifications, from 30 to 40. The concentration of so many languages in such a small territory is indeed remarkable. There are about 8 million speakers of Caucasian languages; their language communities range in size from only a few hundred people to large national groups of millions.

Typological organization

The Caucasian languages fall into three typologically well-defined language families: the Northwest Caucasian, or Abkhazo-Adyghian, languages; the Northeast Caucasian, or Nakho-Dagestanian, languages; and the South Caucasian, or Kartvelian, languages (also called Iberian). From the typological point of view, the Northwest and Northeast Caucasian groups present opposite structural types, with South Caucasian holding an intermediary position.

The exact genetic relationships of the Caucasian languages are still unclear on many points, not only in regard to interrelationships of the three major groups but also to some internal groupings. Although the genetic relationships between Northwest and Northeast Caucasian seem probable, the interrelationships of North and South Caucasian are as yet uncertain because of the absence of any regular sound correspondences between them. At the present stage of comparative Caucasian linguistics, North Caucasian and South Caucasian must be viewed as separate language families.

The theories relating Caucasian with such languages as Basque and the non-Indo-European and non-Semitic languages of the ancient Middle East also lack sufficient evidence and must be considered as inconclusive.

KARTVELIAN (SOUTH CAUCASIAN) LANGUAGES

Languages of the group. The Kartvelian (South Caucasian) language family comprises Georgian, Mingrelian (Megrelian), Laz (or Chan), and Svan. The speakers of these languages constitute the Georgian nation and numbered 4.2 million in the mid-1990s.

Georgian. Georgian (self-designation: *kartuli ena*), used as the language of literature and instruction, is the state language of the Republic of Georgia. It is common to all speakers of the Kartvelian languages within Georgia. Beyond the borders of Georgia, Georgian is spoken in the adjacent regions of Azerbaijan and northeastern Turkey. There are also 14 villages of Georgian speakers in the province of Eṣfahān, Iran.

The designation Georgian that is used in the European languages was coined during the Crusades; it is based on Persian *gorji* (Georgian), from which the Russian *gruzin* was also derived. The Greek term *íbēres* (Georgians) is connected with an Old Iranian name for Georgia.

The dialects of Georgian fall into two groups—East and West Georgian—divided by the Suram Mountains. These exhibit only slight differences.

Georgian literary tradition

Among the Caucasian languages, only Georgian has an ancient literary tradition, which dates back to the 5th century AD, when the oldest datable monuments were inscribed in an original script. With regard to the order of the alphabet and the shape of some characters, this Old Georgian script is presumed to have been derived from the Greek alphabet. The modern Georgian writing system is based on the round-form cursive, which was developed from the angular book script of the 9th century; the latter was a direct descendant of the Old Georgian script. The Georgian writing system includes a symbol for each of the distinctive sounds (phonemes) of the Georgian language.

During the Old Georgian period (from the 5th to the 11th century), original and translated literary monuments were produced, among them the Georgian translation of the Bible. The conventions of the New Georgian literary language, ultimately established in the mid-19th century on the basis of an East Georgian dialect, originated in the secular literature of the 12th century. New Georgian differs structurally in many respects from Old Georgian, but the old language is still comprehensible to the Georgians of today. Until the beginning of the 19th century, Old Georgian was still in use in religious services and theological writings.

Mingrelian. The Mingrelian language (self-designation: *margaluri nina*) is spoken in the territory north of the Rioni River and west of the Tskhenis-Tskali River and along the Black Sea coast from the mouth of the Rioni up to the city of Ochamchire. The language is unwritten; Georgian is used as a literary language.

Laz. The Laz language (self-designation: *lazuri nena*) is spoken along the Black Sea coast from the Chorokh River (Georgia) to south of Pazar (Atina) in Turkish territory. The language is unwritten, Georgian being used as the literary language in Georgia and Turkish in Turkey. In view of the structural closeness between Mingrelian and Laz, they are sometimes considered as dialects of a single language.

Svan. The Svan language (self-designation: *lušnu nin*), also unwritten, is located south of Mount Elbrus, in the high valleys of the upper Tskhenis-Tskali and its tributary Kheledula and in the valleys of the upper Inguri River. There are four fairly distinct dialects: Upper and Lower Bal in the Inguri region, and Lashkh and Lentekh in the Tskhenis-Tskali region. Georgian and Russian are used as literary languages.

Linguistic characteristics. Correspondences between sounds and meanings in words and word elements provide a basis for considering the Kartvelian languages as being closely related and descended from a common ancestral language (a protolanguage).

Phonology. The sound system of the Kartvelian languages is relatively uniform, with only the vowel systems exhibiting considerable differences. Apart from the five cardinal vowels *a, e, i, o, u,* which exist in all the Kartvelian languages, the Svan dialects show several additional vowels: the front (or palatalized) vowels, *ä, ö, ü,* and a high central vowel, *ə* (as the *a* in English "sofa"). All these vowels also have distinct lengthened counterparts, thus giving a total of 18 distinctive vowels in some dialects of Svan. Vowel length is not distinctive in the other Kartvelian languages.

Varieties of Kartvelian consonants

Within the Kartvelian consonant system the stops and affricates have voiced, voiceless, and glottalized varieties. (Stops are produced by complete but momentary stoppage of the breath stream some place in the vocal tract; affricates are sounds begun as stops but released with local friction, such as the *ch* sounds in "church." Voiced sounds are made with vibrating vocal cords; in voiceless sounds, the vocal cords do not vibrate; glottalized consonants, indicated in phonetic transcription by dots below or above certain letters, are pronounced with an accompanying closure of the glottis [the space between the vocal cords].) Fricative sounds (*e.g., s, z, v*), which are characterized by local friction, have only voiced and voiceless types.

Although most word roots begin with one or two consonants, instances of long consonant clusters in word-initial position occur quite frequently, especially in Georgian, in which such clusters may comprise up to six consonants—*e.g.,* Georgian *prckvna* "peeling," *msxverp̣li* "sacrifice," *brʒola* "fighting."

Grammatical characteristics. The Kartvelian languages exhibit a developed system of word inflection (*e.g.,* the use of endings, such as English "dish, dishes" or "walk, walks, walked") and derivation (word formation). Derivation is characterized by compounding, the combination of words to form new words, as well as by affixation, the addition of prefixes and suffixes—*e.g.,* Georgian *kartvel-i* "Georgian," *sa-kartvel-o* "Georgia"; Mingrelian *žir-i* "two," *ma-žir-a* "second."

The verb system distinguishes the categories of person, number (singular and plural, with differentiation of inclusive and exclusive plural in Svan), tense, aspect, mood, voice, causative, and version (the latter defines the subject–object relations). These categories are expressed mainly by the use of prefixes and suffixes, as well as by internal inflection (changes within the verb stem), which is frequently a redundant grammatical feature.

The system of verb conjugation in Kartvelian languages is multipersonal; that is, the verb forms can indicate the person of the subject (the agent) and of the direct or indirect object by the use of special prefixes. (The subject of the third person is marked by endings in Georgian and Mingrelo-Laz and by a lack of ending in Svan.) An example is Georgian *m-ç̣er-s* "he writes to me," *m-xaṭav-s* "he paints me," in which *m* denotes the first person as object and *s* marks the third person as subject. The finite verb forms fall into three series of tenses: the present tense, the aorist (indicating occurrence, usually past, without reference to completion, duration, or repetition), and the perfect or resultative (denoting an action in the past not witnessed by the speaker).

There is a developed system of preverbs, elements preceding the verb stem and attached to it, with local meaning indicating location of the action in space, as well as its direction (especially in Mingrelian and Laz). Simple preverbs are combined into complex ones. The preverbs are also used to mark the aspect (nature of the action indicated by the verb, with reference to its beginning, duration, completion), which is used for the formation of future and aorist forms—*e.g.,* Georgian *ç̣er-s* "he writes" versus *da-ç̣er-s* "he will write" and *da-ç̣er-a* "he wrote."

The Kartvelian nominal system

The nominal (noun, pronoun, adjective) system is distinguished by less structural complexity than the verb system and has cases varying in number from 6 to 11. The six cases common to all the Kartvelian languages are: nominative, marking subject of the intransitive verb; ergative (see below), modified in Mingrelian and Laz; genitive, marking possession; dative, marking indirect objects; ablative–instrumental, expressing relations of separation and source and means or agency; and adverbial, expressing goal of the action—*e.g.,* "to make it." There are also some secondary local cases (in New Georgian, Mingrelian) that indicate location and direction toward the object as well as from the object (rendered in English by such prepositions as "in," "on," "to," "from," and so on). The nominal system does not distinguish gender, and there are no special articles (such as English "a," "the").

A basic feature of Kartvelian syntax is the ergative construction of the sentence. The subject of a transitive verb (the agent) is marked by a special agentive, or ergative, case, while the case of the direct object is the same as that of the subject with intransitive verbs, traditionally called the nominative case—*e.g.,* Georgian *ḳac-i* (nominative) *midis* and Svan *māre* (nominative) *esγri,* "the, or a, man goes" but Georgian *ḳac-ma* (ergative) *moḳla datv-i* (nominative) and Svan *mārēm* (ergative) *adgär däšdw* (nominative) "the, or a, man killed the, or a, bear." A specific feature of the Georgian and Svan ergative construction is its restriction to the aorist series (*i.e.,* that showing simply occurrence). In the present-tense series the subject (agent) of transitive as well as intransitive verbs is put into the same nominative case, and the direct object is in the dative—*e.g.,* Georgian *ḳac-ma* (ergative) *moḳla datv-i* (nominative) "the man killed a bear" (aorist), but *ḳac-i* (nominative) *ḳlavs datv-s* (dative) "the, or a, man kills the, or a, bear" (present tense). In Mingrelian the ergative case with the formative (suffix) *-k* extends in the aorist series to the constructions with intransitive verbs and results in a formation of two distinct subject cases. In Laz, conversely, the case with the formative *-k* extends to the constructions with transitive verbs in the present-tense series.

Vocabulary. The genetic closeness of the Kartvelian languages is evidenced by a large number of structural correspondences and of common lexical (vocabulary) and grammatical items. Though the Kartvelian languages abound in ancient loanwords from Iranian, Greek, Arabic, Turkish, and other languages, it is nevertheless possible to single out the basic vocabulary and grammatical elements of original Caucasian origin, which exhibit a system of regular sound correspondences. The common Kartvelian vocabulary comprises the kinship terms, names of animals, birds, trees, and plants, and parts of the body, as well as different human activities, qualities, and states. The words that are used for the numerals from 1 to 10 and the word for "hundred" are also original common Kartvelian terms.

Ablaut in Proto-Kartvelian

Proto-Kartvelian. A comparative study of the Kartvelian languages enables specialists to outline the general structure of the parent language, called Proto-Kartvelian, which yielded the known Kartvelian, or South Caucasian, languages. One of the most characteristic features of the Proto-Kartvelian language is the functional vowel alternation, or ablaut; different forms of a word root or word element appear either with a vowel (**e, *a, *o*), called full grade, or without a vowel, called zero grade. (An asterisk [*] indicates that the following form is not attested but has been reconstructed as a hypothetical ancestral form.) In a sequence of word elements (called morphemes) only one element may occur in full grade, the others being in either zero or reduced grade forms (*i.e.,* in a form with **i*). To a word root with a full-grade vowel, for example, a suffix in zero may be added, and vice versa: **der-ḳ-* (intransitive) "stoop, recline" and **dr-eḳ-* (transitive) "bend." When a full-grade ending is added to these stems, the preceding full-grade element is shifted to zero or a reduced grade; *e.g., *der-ḳ-* plus the ending **-a* becomes **dr̥-ḳ-a.* In such patterns the lengthened grade, a long vowel, may also appear.

These ablaut patterns, strikingly parallel to those of the Indo-European languages, and other linguistic features may have arisen in Proto-Kartvelian as a result of contacts with Indo-European at a comparatively early date. Such contacts between Kartvelian and Indo-European are further evidenced by a number of Indo-European loanwords in Proto-Kartvelian, such as Proto-Kartvelian **ṭep* "warm" (compare Indo-European **tep* "warm"), Proto-Kartvelian **m̥ḳerd* "breast" (compare Indo-European **k̑erd* "heart"), and others.

In Mingrelo-Laz the ancient ablaut patterns were eliminated and new forms were set up with a stable, noninterchanging vowel in each word element. The ancient ablauting models were better preserved in Georgian and especially in Svan, in which new ablauting patterns, in addition to the old structures, were established.

The pronominal system of Proto-Kartvelian is characterized by the category of inclusive–exclusive (*i.e.,* there are two forms of the pronoun "we," one including the hearer, and the other excluding him), which survived in Svan but has been lost in other languages of the family. Svan also has preserved a certain number of archaic structural features of the Proto-Kartvelian epoch, further setting it apart from Georgian and Mingrelo-Laz, which share a number of common lexical and grammatical innovations. These features provide evidence that Svan was separated fairly early from the rest of Proto-Kartvelian, which later yielded the Mingrelo-Laz and Georgian languages. (Th.V.G.)

NORTH CAUCASIAN LANGUAGES

The North Caucasian languages are divided into two groups: Abkhazo-Adyghian, or the Northwest Caucasian, languages, and Nakho-Dagestanian, or the Northeast Caucasian, languages.

Abkhazo-Adyghian languages. The Abkhazo-Adyghian group consists of the Abkhaz, Abaza, Adyghian, Kabardian, and Ubykh languages. Adyghians and Kabardians are often considered members of a larger, Circassian group. Abkhaz, with about 90,000 speakers, is spoken in Abkhazia (the southern slopes of the western Great Caucasus, Georgia). The other languages are spread over the northern slopes of the western Great Caucasus. Abazians, who numbered some 20,000 in the Soviet census of 1989, live in Karachay-Cherkessia; Adyghians (120,000), in Adygea; Kabardians (380,000) dwell mainly in Kabardino-Balkaria. Both Adyghians and Kabardians call themselves *adəge.* The Ubykh language, now extinct, was formerly found to the north of the area where Abkhaz is spoken, in the vicinity of Tuapse, Russia. In 1864 Ubykhians as well as a substantial part of the Abkhaz- and Adyghe-speaking population migrated to Turkey, where before long they lost their native tongue. The total number of people speaking Abkhazo-Adyghian languages is about 610,000. Many speakers of Abkhazo-Adyghian languages live in the countries of the Middle East—Turkey, Syria, Jordan, and Iraq.

The extinct Ubykh language

All Abkhazo-Adyghian languages, with the exception of Ubykh, are written. From the dialectological point of view, the Abkhazo-Adyghian languages are not widely differentiated, the differences being mainly of phonetic character. In Abkhaz two dialects are distinguished; Adyghian and Kabardian differentiate four dialects each. Abkhaz and Abaza are very close to each other and are considered by some scholars to be dialects of the same language. The same kind of affinity exists between Adyghian and Kabardian. Ubykh occupies an intermediate position between the Abkhaz-Abaza and Adyghe-Kabardian languages.

Phonology. A characteristic feature of the sound system of the Abkhazo-Adyghian languages is a rather limited number of distinctive vowels—*a* and *ə* (pronounced as the *a* in English "sofa"). Some scholars consider it possible to posit only one vowel, which, depending on the position, can be realized in different ways: *a, ə, i, o, e.* On the other hand, the languages are notable for a great diversity in their consonant systems. The number of consonants distinguished reaches about 70 (in the Abkhaz and Adyghian languages) or even 80 (Ubykh). Along with the consonants that occur in all the Caucasian languages, the Abkhazo-Adyghian languages are characterized by different sets of labialized consonants (formed by rounding the lips), strong (hard or tense) consonants, half-hushing consonants, and velarized consonants (formed with the back of the tongue approaching the soft palate).

Grammatical characteristics. The grammatical characteristics of the Abkhazo-Adyghian languages include an extremely simple noun system and a relatively complicated system of verb conjugation. There are no grammatical cases in Abkhaz and Abaza, and in the other languages only two principal cases occur: a direct case (nominative) and an oblique case, combining the functions of several cases—ergative, genitive, dative, and instrumental. In nouns, possession is expressed by means of pronominal prefixes—*e.g.,* Abkhaz *sarra s-č:ə* "my horse" (literally: "I my-horse"), *wara u-č:ə* "your horse" (pertaining to a man), *bara b-č:ə* "your horse" (pertaining to a woman), and so forth. (The colon [:] indicates that the preceding consonant is a strong consonant.)

The Abkhaz and Abaza languages distinguish the grammatical classes of person and thing (the latter class includes all nouns denoting nonhuman objects). The class of person also differentiates between the subclasses of masculine and feminine.

Verb categories

The verb in the Abkhazo-Adyghian languages has a pronounced polysynthetic character; that is, various words combine to form a composite word that expresses a complete statement or sentence. The most important verbal categories are expressed by prefixes, although suffixes also form tenses and moods. The principal verb categories are dynamic versus static, transitivity, person, number, class, tense, mood, negation, causative, version, and potentiality. "Dynamic versus static" is a verb form expressing action versus state of being; "version" is a verb category denoting for whom the action is intended (compare Georgian *v-çer* "I write," but *v-u-çer* "I write for him"); "potentiality" is a category expressing the possibility of an action (*e.g.,* Abkhaz *s-zə-ɯuam* "I cannot write"). The verb is multipersonal and can denote up to four persons.

Adverbial relationships (such as "where," "when," "how") are expressed by prefixes following the personal markers. On the whole, the verb forms appear as a long string of word elements expressing the above-mentioned categories—*e.g.,* Abkhaz *i-u-z-d-aa-sə-r-g-an* "that (thing)-you (masculine)-for-them-hither-I shall-make-bring" (*i.e.,* "I shall make them bring that for you"). In a sequence of prefixes, up to nine morphemes are possible.

The simple sentence has three constructions: indefinite, nominative, and ergative (in Abkhaz and Abaza only indefinite). An indefinite construction has the subject in the indefinite case (*i.e.,* not marked with a special suffix); a nominative construction has the subject in the nominative case. The same personal markers, depending on their arrangement, can denote both the subject and various objects—*e.g.,* Abkhaz, *wara sara u-s-šwejṭ* "I kill you (masculine)," *sara wara s-u-šweiṭ* "you (masculine) kill me."

Nakho-Dagestanian languages. The Nakho-Dagestanian group consists of the Nakh and Dagestanian languages. Some investigators subdivide the Nakho-Dagestanian languages into two independent groups: Central Caucasian languages (Nakh) and East Caucasian languages (Dagestan), although the great proximity of these groups, and their equal remoteness from the Abkhazo-Adyghian languages, may justify regarding them as a common group of languages.

The Nakh languages consist of Chechen (890,000 speakers), Ingush (210,000), and Bats (or Tsova-Tushian, about 3,000 speakers). The Chechens and Ingush live in Chechenia and Ingushetia; the Bats dwell in the village Zemo-Alvani in the Akhmeta district of northeastern Georgia. Both Chechen and Ingush, which are fairly similar to one another, are written. Bats speakers, whose language is not written, use Georgian as their literary language.

The Dagestan languages are numerous. The following groups can be distinguished:

The Avar-Andi-Dido languages. These occupy the central and western part of Dagestan and part of the Zakataly region in northwestern Azerbaijan. The member languages are the Avar language; the Andi subgroup of languages, including Andi, Botlikh, Godoberi, Chamalal, Bagvalal, Tindi, Karata, and Akhvakh; and the Dido subgroup, including Dido (Tsez), Khvarshi, Hinukh, Bezhta, and Hunzib.

Avar as lingua franca

Of these tongues, the language with the most speakers (about 530,000) is Avar, which has literary status. None of the Andi-Dido languages are written; Avar is used as the literary language. Most of them are spoken by fewer than 10,000 people. From ancient times the Andi-Dido nationalities have used the Avar language for intertribal communication. Avar is still widely known and spoken among them. The Andi languages are phonetically and grammatically very close to each other. The same affinity is observed among the Dido languages. In respect to dialectology, the majority of Avar-Andi-Dido languages are widely differentiated.

The Lak-Dargin languages. Lak (also spelled Lakk, with some 100,000 speakers) and Dargin (or Dargwa, with 350,000) are spoken in the central part of Dagestan. Both are written languages. The Lak language is quite homogeneous with regard to its dialects; Dargin, however, possesses several diversified dialects—sometimes considered as separate languages (*e.g.,* Kubachi). Some view Lak and Dargin as independent language groups.

The Lezgian languages. This language group includes Lezgi (with 240,000 speakers in Dagestan and about 170,000 in Azerbaijan); Tabasaran (about 90,000); Agul (about 12,000); Rutul (about 15,000); Tsakhur (about 11,000); Archi (fewer than 1,000); Kryz (about 6,000); Budukh (about 2,000); Khinalug (about 1,500); and Udi (about 3,700). The majority of Lezgi languages are spoken in southern Dagestan, but some of them (Kryz, Budukh, Khinalug, Udi) are spoken chiefly in Azerbaijan; and one village of Udi speakers is located in Georgia. It is important to note that in Azerbaijan, as well as earlier in Russia, all Dagestanians—including Avars—referred to themselves as Lezginians. Among the Lezgian languages, only Lezgi and Tabasaran are written. Archi, Khinalug, and Udi are the most divergent languages of the Lezgian division. The Udi language is believed to be one of the languages of ancient Caucasian Albania.

Linguistic characteristics. *Phonology.* The sound systems of the Nakho-Dagestanian languages are diverse. There are up to five vowels (*a, e, i, o, u*); in some languages *o* is only now becoming an independent distinctive unit. Along with these cardinal vowels, in a number of languages there are also long and nasalized vowels (the Andi languages), pharyngealized vowels (in Udi), and labialized vowels (in Dido). In the Nakh languages (such as Chechen) the vowel system is fairly intricate, the number of distinctive vowels amounting to 30 (including diphthongs and triphthongs).

The consonant systems of the Nakh languages are relatively simple, coinciding, on the whole, with those of the South Caucasian languages (apart from a number of pharyngeal consonants characteristic of all the Nakh languages and a lateral sound peculiar to Bats).

The opposition of strong and weak voiceless consonants is typical of the majority of the Dagestanian languages. This contrast has been lost in a number of languages and dialects—for example, in the Dido languages and in some dialects of Avar. The labialized clusters *kw, qw, sw,* and so on, are widespread. In the Avar-Andi-Dido languages and in Archi there are fricative and affricate lateral sounds (*i.e.,* different types of *l*), with the maximum possible number being six (in Akhvakh).

Types of stops in Caucasian

All the Caucasian languages have a series of stops of three types—voiced, voiceless aspirated, and glottalized (*i.e.,* pronounced, respectively, with vibrating vocal cords; with vocal cords not vibrating but with an accompanying audible puff of breath; and with accompanying closure of the glottis [space between the vocal cords]). In some languages strong and weak consonants also contrast. Usually, in the languages with a strongly developed vowel system, the system of consonants is comparatively simple (*e.g.,* Chechen, Ingush, Dido), and vice versa (*e.g.,* Avar, Lak, and Dargin have complicated consonantisms and relatively simple vowel systems).

Grammatical characteristics. There are several common structural features in morphology (word structure), the most characteristic being the existence of the grammatical category of classes (eight classes in Bats; six in Chechen and Andi; five in Chamalal; four in Lak; three in Avar; two in Tabasaran).

In a number of languages (Lezgi, Udi) noun differentiation by classes has disappeared. The class of "thing" is distinguished from the "person" class, which can be differentiated into the subclasses of masculine and feminine. Compare, for example, Avar *emen w-aç̌ana hani-w-e* "father has come here" (in which *w* is equivalent to the marker of the class of masculine person), *ebel j-aç̌ana hani-j-e* "mother has come here" (in which *j* is equivalent to the marker of the class of feminine person), and *č̣u (kaɣat) b-aç̌ana hani-b-e* "a horse (a letter) has come here" (in which *b* is equivalent to the marker of the class of thing). In the plural there are usually fewer grammatical classes denoted.

Nouns have many cases, both in singular and in plural; there are cardinal cases (nominative, ergative, genitive, dative) and local cases that denote the location of a thing ("on," "in," "near," "under"), with a specification of movement ("where," "which way," "from where," "over what"). The ergative case, the case of the real subject of transitive verbs, is present in all the Nakho-Dagestanian languages. Nouns have different stem forms in the nominative and the oblique (non-nominative) cases—*e.g.,* Avar *gamaç̌* "a stone" (nominative), *ganç̌-i-c:a* (ergative), and *ganç̌-i-da* "on the stone." In pronouns the category of inclusive–exclusive is distinguished—*e.g.,* Avar *nil* "we with you," *niž* "we without you."

The class of the noun in the nominative case (*i.e.,* in the case of the subject of intransitive verbs and of the direct object of transitive verbs) is reflected in the verb—*e.g.,* Avar: *was* (nominative, class I) *w-aç̌ana* "the boy has come," *jas* (nominative, class II) *j-aç̌ana* "the girl has come."

In the Lezgi language, a characteristic structural feature is agglutination, the combination of various elements of distinct meaning into a single word. A typical feature of Nakho-Dagestanian syntax is the presence of the ergative construction of the sentence (the subject of transitive verbs is put in the ergative case and the real object in the nominative case). Complex sentences are usually formed with participial and adverbial–participial construction; *e.g.,* Avar *haniwe waç̌araw či dir wac: wugo* "the man who arrived here is my brother" (literally, "the here arrived man my brother is").

Sources of loanwords

Vocabulary. The original vocabulary of the North Caucasian languages has been fairly well preserved in the modern languages, although many words have been borrowed from Arabic (through Islām), the Turkic languages, and Persian. There are also loanwords that have been taken from the neighbouring languages (Georgian, Ossetic). Russian, which was a major influence from the late 19th century, was for decades the main source for new words, especially technical terminology.

Writing and alphabets. The written languages of the area are the state languages. Newspapers, magazines, and books, as well as radio and television programs, use the local languages, and children in primary schools are taught in their mother tongue.

The alphabets of the North Caucasian written languages (Abkhaz, Abaza, Adyghian, Kabardian, Chechen, Ingush, Avar, Lak, Dargin, Lezgian, Tabasaran) are based on the Cyrillic alphabet, which was introduced for these languages in 1936–38 (in Abkhaz, from 1954). Previously, from 1928, the modified Latin alphabet was used; it superseded the Arabic script, which was adapted to the local languages in the Soviet period, at a time when a number of North Caucasian languages became literary languages.

Earlier attempts had been made to provide written forms of some Caucasian languages. In the 18th century an insignificant number of monuments were created (with the use of Arabic writing) in Lak and Avar. Stone crosses with Old Georgian–Avar bilingual inscriptions, dating from not later than the 14th century AD, have been preserved in central Dagestan. (T.E.G.)

AFRO-ASIATIC (HAMITO-SEMITIC) LANGUAGES

The Hamito-Semitic languages, a family of genetically related languages, developed from a common parent language that presumably existed about the 6th–8th millennia BC and was perhaps located in the present-day Sahara. Also known as the Semito-Hamitic, Erythraean, Afro-Asiatic, and Afrasian language group, it is the main language family of northern Africa and southwestern Asia and includes such languages as Arabic, Hebrew, Amharic, and Hausa. The total number of speakers is estimated to be more than 200,000,000.

The term Hamito-Semitic, or Semito-Hamitic, was introduced by a German Egyptologist, Karl Richard Lepsius, in the 1860s. Although it has become traditional, it is an unfortunate label in suggesting that the family is divided into a group of Semitic and a group of Hamitic languages; in fact, the family has at least four other branches of the same order as the Semitic languages. The term Erythraean is inappropriate in implying that the family originated on both shores of the Red Sea, an assumption that cannot be proved; and Afro-Asiatic (proposed by an American linguist, Joseph Greenberg, in 1950) may be too comprehensive insofar as it suggests that all the languages of Africa and Asia are included. Igor Diakonoff, a Russian linguist, has suggested the term Afrasian, meaning "half African, half Asiatic," which corresponds to the area of the actual distribution of the languages of this family since at least the 5th millennium BC.

The five branches of Hamito-Semitic

The languages belonging to the Hamito-Semitic family can apparently be subdivided into branches representing dialects of the original parent language—namely, Semitic, Egyptian, Berber, Cushitic, and Chadic. Some linguists deny the genetic affinity of the Chadic languages with the other branches of Hamito-Semitic, while others (*e.g.,* Joseph Greenberg) accept it. Certain scholars have expressed doubts concerning the Hamito-Semitic character of some of the Chadic languages but not of others. Among the linguists who classify the Chadic languages as Hamito-Semitic there is some hesitation as to the degree and character of their affinity with the languages of the Cushitic branch, especially with West Cushitic. On the basis of the low percentage of vocabulary items held in common between the West Cushitic languages and the other Cushitic languages, some scholars classify West Cushitic as a separate branch of Hamito-Semitic, called Omotic. There is, however, a probability that the parent language common to Omotic and the Cushitic languages proper is not the Common Hamito-Semitic protolanguage but a later dialect (namely, Common Cushitic) and that Omotic (West Cushitic) is thus, nevertheless, a subgroup of Cushitic. Others connect Omotic with the Chadic group.

Some linguists have suggested that the Hamito-Semitic languages are related to the Indo-European languages; others have favoured the existence of a superfamily, including the Hamito-Semitic, Indo-European, Altaic, Finno-Ugric (Uralic), Kartvelian, and Dravidian languages; but most scholars regard such far-flung genetic ties as unproven and, indeed, hardly provable.

Because there has been a considerable difference of opinion as to the criteria to be applied when identifying a language as Hamito-Semitic, the basic principles of linguistic classification as applicable in this case should be stated. The only real criterion for classifying certain languages together as a family is the common origin of their most ancient vocabulary as well as of the word elements used to express grammatical relations. A common source language is revealed by a comparison of words from the supposedly related languages expressing notions common to all human cultures (and therefore not as a rule likely to have been borrowed from a group speaking another language) and also by a comparison of the inflectional forms (for tense, voice, case, or whatever).

Reconstruction of Common Hamito-Semitic

If, as a result of a step-by-step reconstruction of forms having existed at earlier periods, scholars arrive at an identical original phonological structure for each of the words or word elements compared in several different known languages, then such original forms can be ascribed to a common language, which, in the case of the languages here discussed, is conventionally termed Common Hamito-Semitic (or Proto-Hamito-Semitic). It also stands to reason that wherever one parent language has existed the daughter languages must to some degree reflect some of its grammatical characteristics.

Despite the work of several scholars, only an approximate and provisional reconstruction of the parent language forms of Hamito-Semitic has so far been made. More work, however, has been done in comparing the language typologies.

COMMON CHARACTERISTICS

Certain typological features seem to have been common to all Hamito-Semitic languages at an early stage of their development. Among the phonological features are (1) a six-vowel system (*a, i, u, ā, ī, ū*—that is, short and long *a, i, u*), perhaps developed from an earlier two-vowel system (of **a* and **ə* [pronounced as the *a* in "sofa"]; an asterisk before a sound or a word-form indicates that it is not attested but is reconstructed hypothetically); (2) pharyngeal fricative consonants, indicated by the symbols ʿ (voiced) and *ḥ* (voiceless) and produced in the region of the pharynx; (3) the functioning of the glottal stop (articulated by closing the glottis, the space between the vocal cords) as a separate distinctive sound (phoneme)—this is conventionally indicated by ʾ; (4) the use of the semivowels *u̯* (*w*) and *i̯* (*y*) in the structural role of consonants; and (5) three types of consonants: voiceless, voiced, and "emphatic," the last type being phonetically realized either as voiceless consonants combined with a glottal stop, as pharyngealized voiceless or voiced consonants, or as consonants in which the air is drawn into the mouth (injective [preglottalized], or implosive), consonants in which the tongue is retracted from the usual position (velarized), or in which the tongue tip is curled upward toward the hard palate (retroflex or cerebral).

Common morphological features include (1) word bases for verbs and for nouns derived from verbs consisting of two elements that interweave with one another, a "root" consisting of consonants, and a "scheme" consisting of vowels (for examples see below); (2) a predominance of word roots consisting of three consonants over roots of two consonants; (3) a strongly developed system of infixation—*i.e.,* the insertion of elements within the root of a word to show grammatical changes and form new words with related meaning; and (4) a comparatively poorly developed system of prefixes and suffixes.

Morphological similarities among the languages

In the area of morphological typology, there are numerous similarities among the Hamito-Semitic languages, such as a system of declension of the noun and pronoun with at least three cases (nominative, genitive, accusative, with

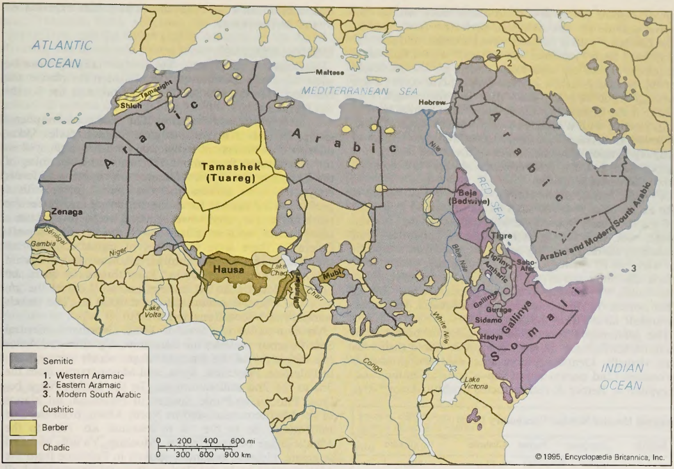


Figure 27: Distribution of the modern Hamito-Semitic languages.

traces of a still earlier system including only the agentive [ergative], and unmarked [zero] cases, or agentive, genitive, and unmarked). There are three numbers in the noun, pronoun, and verb—singular, dual, and plural. An event considered from the point of view of the resulting state, as opposed to the point of view of the action itself, is expressed by a special predicative (zero) form of the noun that later developed into a new verbal "tense." In addition, there is a well-developed binary system of verbal aspects, indicating the mode of an action (*i.e.*, punctual contrasts with durative, or perfective [completed action] contrasts with imperfective [ongoing action]), but tenses and voices of the verb remained undeveloped until the later stages. Pronominal possession markers and object markers in the form of suffixes are another common Hamito-Semitic feature, as are the prefixing of certain actor markers to the verb and a two-gender system in the noun, pronoun, and verb, perhaps developed from a still earlier system of many genders. In syntax, the Hamito-Semitic languages show certain favoured types of attributive constructions, among other common characteristics.

The above inherited Hamito-Semitic characteristics are listed, for each linguistic level, in the approximate reverse order of their stability. Languages retaining all or most of these features can be classified as belonging to the Ancient Stage of Hamito-Semitic; those that retain no less than two-thirds of the ancient consonantal system and about one-half to two-thirds of the above-listed other features belong to the Middle Stage; those that have lost more than half of these characteristics belong to the New Stage. At the New Stage, however, there are usually enough of these features still preserved to identify the language as belonging to the Hamito-Semitic family, and most of the other features can, as a rule, be reliably reconstructed for one of the former stages of its development. Moreover, the original form of the word elements that express the typical Hamito-Semitic grammatical features is usually apparent in all languages of the family. All modern Hamito-Semitic languages except Literary Arabic and Hebrew belong to the New Stage.

The character of the relationship between the five branches of the Hamito-Semitic family—Semitic, Egyptian, Berber, Cushitic, and Chadic—can best be seen by comparing their systems of verbs and pronouns. There are several types of verbal systems in Hamito-Semitic, but all of them (with the exception of the Egyptian, which has developed in a quite different direction) can apparently be traced back to one single system. In this system the action (including intransitive action) is expressed by a verbal form proper, with a prefixed actor marker (singular: 1st person *'*a*-, 2nd **ta*-, 3rd **ya*-) probably deriving from a separate personal pronoun in an oblique case; the state is expressed by a form of a noun used as a predicate, plus a personal pronoun in the direct case (this is called stative). Hamito-Semitic apparently developed from a protolanguage with an ergative type of sentence construction (in which there is a special case denoting the agent of an action but no marker for the subject of a state and the direct object of an action) to a language of the nominative type (in which the subject both of an action and a state is always in the nominative case and the direct object is in the accusative case). At the same time, the predicate of state (the so-called stative) developed into either a perfective aspect (marking completion of the action of the verb) or a past tense of the verb, or it disappeared altogether. There are, however, enough traces of its existence in all branches of the family (*e.g.*, in Egyptian, in Kabyle of the Berber branch, in Sidamo of the Cushitic branch, in Mubi of the Chadic branch, and in all Semitic languages) to see that the form goes back to the parent language.

Hamito-Semitic verbal systems

As for the verbal forms that express action and have a prefixed actor marker, there is some discrepancy of opinion. Some scholars posit for the parent language only one form. It may be, however, that there were two forms for the transitive, a perfective and an imperfective type, and possibly only one form for the intransitive type.

In several languages of the New Stage, new verbal types have developed for all aspects and tenses, particularly in the languages of the Cushitic branch (the Northern, Eastern, and Central groups, in part; and the Southern and Western [Omotic] group, always), the Chadic branch (in most languages), and the Semitic branch (typically in Neo-Syriac). These verbal forms consisted originally of a noun (for the most part, derived from a verb) plus an auxiliary verb with a prefixed actor marker. Everywhere, as a rule, the perfective aspect (or the past tense) is formed from bases of the auxiliary verb with a reduced vowel scheme in the verbal base, while the imperfective aspect (or the present/future tense) is formed from bases with a full vowel scheme (*cf.* the Akkadian perfective form **yaprus* "he divided," with a reduced vowel scheme, and the imperfective form **yaparras* "he divides," with a full vowel scheme). (There are also forms based on the participle of the auxiliary verb; *e.g.*, Neo-Syriac *biktā-vōvin* "I am writing" from **bi-ktābā-hāwē-ʾănā* "in-write-being-I.")

In that the Central Semitic verbal system (which has the imperfective with a reduced vowel scheme, as in Arabic, Hebrew, and Aramaic) is restricted to only two groups of languages inside only one branch of the entire family, it is improbable that it is this verbal system that is descended from the parent language.

Stem modification

A typical feature of the Hamito-Semitic verbal system is the existence of so-called stem modifications—*i.e.*, groups of systematically related verbal stems deriving from a single root, each having its own type of semantics—that variously characterizes the action or state from the point of view of its quality, quantity, frequency, causal relations, direction, and so on. In Hebrew, for example, *šābar* "he broke," *šibbēr* "he broke to pieces," *hišbīr* "he let (him) break out," and *nišbar* "he was broken, destroyed, stranded" all are from the root *šbr*.

The pronominal systems in the different branches of Hamito-Semitic are more or less alike. Some pronouns are virtually identical everywhere; *e.g.*, the possessive pronouns (2nd person masculine—"your": Semitic **-ka*, Egyptian *-k*, Berber *-k*, reconstructed Cushitic *-ka* or **-k*^{w}*a*, Chadic [Hausa] *-ka*). Suffixed pronouns expressing the object of the verb are very similar to the possessive.

Table 48: Common Hamito-Semitic Vocabulary Items

	Semitic	Egyptian	Berber	Cushitic	Chadic
"bone"	*qōṣ* ("thorn," Hebrew)	*qs*	*i-ghəs*	**m-kkac*	**ḳasi* (Hausa)
"to die"	**mūt*	*m(w)t*	*əmmət*		*mutu* (Hausa)
"dog"	*kal-b-*			**kala-kara-*	**kala-kara-*
"eye"		*ir. t*		**yil*	**yil*
"heart"	**libb-*	*ib*		**libb-*	
"I"	*ʾanāku* *ʾanā* *ʾanī*	*ink* *anok* (Coptic)	*n(ə)ki*	**(ʾ)ani* **ta-(ʾ)ani* *anu* *ana*	*an* (Sura) *ni* (Hausa) *n-ani* (Kanakura)
"jackal" ("wolf," "dog")		*wnš*	*uššən*	**waẑẑ* *(from *wanẑ-*?) (Highland Eastern Cushitic)	
"man"	**mut-*	**mt*			*mito* (Jegu) *miji* (Hausa)
"name"	**šim-*		*i-səm*	**sim-*	**sim(m)-* *summ-* *sūn-*
"thou"	*ʾanta* *ʾatta*	*ntk* *əntok* (Coptic)		**ʾatta*	
"tongue"	**laš-ān-*	*ns* (pronounced *las*)	*i-ls*		**ha-ls(e)* (Hausa)
"tooth"	**šinn-*	*sn* ("harpoon")	*-sin*		**sinn-*(?)
"two"	**ṯhin-*	*sn*	*sin*	**čan-*("two equal parts")	
"water"	**mā-may-*	*m(y)-w*	*a-ma-n*	*mā-n* ("sea") (Somali)	

*Reconstructed form.

Linguistic substrata

The diverging of the branches and the individual languages of the Hamito-Semitic family from the common ancestral language, although mainly explained by the internal development of the languages after loss of contact, also results to a great extent from the influence of different linguistic substrata. Thus, the ancient Hamito-Semitic language had in many cases probably spread to originally alien populations. This view is supported by the different racial types of the speakers. In some cases the substratum language (*i.e.*, that of the original population) can be identified —*e.g.*, Sumerian, Hurrian, and others for North Semitic; Nilo-Saharan and East Sudanese for Cushitic; East Sudanese and possibly some others for Chadic. The least substratum influence seems to have been experienced by the Berber branch.

SEMITIC LANGUAGES

Languages of the group. The Semitic languages can be subdivided into four groups: the Northern Peripheral, the Northern Central, the Southern Central, and the Southern Peripheral.

Akkadian

Northern Peripheral Semitic. The Northern Peripheral group, from the Ancient to Middle Stage, includes Akkadian with its dialects of Babylonian and Assyrian, spoken in Mesopotamia from about 3200 BC to the beginning of the Christian Era. Typical features are stative verb forms conjugated with suffixes and two verbal forms with a prefixed actor marker for the imperfective and perfective (with full and reduced vowel schemes, respectively; later a new "perfect" with an infixed *-ta-* in the stem developed). Originally there were five cases of the noun, plus an unmarked form for the nominal predicate and the noun without grammatical relations. Later three cases remained but were lost in the 1st millennium BC. Loss of the *gh, ḥ, ʿ*, and *h* sounds occurred from *c.* 2000 BC. The vowels were *a, e, i, u* (both long and short).

Hebrew

Northern Central Semitic. The Northern Central Semitic group includes the Canaanite, Ugaritic, and Amorite languages of the Ancient Stage, which were spoken in Palestine, Phoenicia, Syria, and Mesopotamia from the 3rd to the 2nd millennium BC. To the Middle Stage belongs Phoenician-Punic, spoken in Phoenicia, on islands of the Mediterranean, and in North Africa, from the 2nd millennium BC to the 1st millennium AD. Also to the Middle Stage belong Hebrew, Moabite, Ya'ūdī, and Old Aramaic. Hebrew, originally spoken in Palestine from the 13th century BC to the 2nd century AD, later spread all over the world as a written language. At present there are about 2,600,000 Hebrew speakers in Israel. Moabite and its kindred dialects in the Transjordan were alive in the 1st millennium BC but are now extinct. Ya'ūdī, spoken in northern Syria in the 9th century BC, is also extinct. Old Aramaic, from Syria and Mesopotamia, existed from the 14th century BC(?) through the 15th century AD. Its oldest written texts date from the 9th century BC. The dialects of Aramaic include Ancient Aramaic proper; Imperial Aramaic (the official language of Assyria and Achaemenid Persia, including also Biblical Aramaic, or Chaldean); Western Aramaic, with Palmyrenean, Nabataean, Palestinian, Galilean, and other varieties; Eastern Aramaic, including Syriac (Edessan, with subdialects), Babylonian Talmudic, and Mandaic. Most Aramaic dialects gave way to Arabic beginning with the 7th century AD.

Aramaic dialects

The New Stage of Northern Central Semitic is represented by New West Aramaic, or Maʿlūla, in Syria, with a small number of speakers, and Neo-Syriac, or "Assyrian," in Iraq (al-Mawṣil), Turkey (Ṭūr-ʿAbdīn), Iran (Urmia), the Soviet Union, and the United States, with about 200,000 speakers.

Typical features of the Northern Central Semitic group are the perfective aspect with suffix conjugation and the imperfective aspect with prefix conjugation and stems with a reduced vowel scheme. (An entirely new verbal system developed in Neo-Aramaic.) The group is also characterized by the article *ha-* (prefixed, in Hebrew and Punic) or *-ā* (suffixed, in Aramaic). The system of declension was lost from the Middle Stage on. In this group the number of vowel qualities increased beyond just *a, i, u,* while the number of consonants diminished considerably from the Middle Stage on. The sounds *p, t, k, b, d, g* became aspirated after vowels—*i.e.*, pronounced with an accompanying puff of breath (and are now pronounced as *f, t* or *s, kh, v, d, g* in Modern Hebrew).

Arabic

Southern Central Semitic. The Southern Central group includes Classical, or Literary, Arabic belonging to the Ancient and Middle stages. Originally spoken in Arabia, Classical, or Literary, Arabic is now found from the Indian to the Atlantic Ocean and has been attested from the 5th century BC to the present time. From the New Stage come the modern Arabic dialects, some of them mutually unintelligible. They have about 130,000,000 speakers all over northern Africa, on the Arabian Peninsula, in Jordan, Israel, Lebanon, Syria, and Iraq, and in some districts of Turkey, Iran, and the Soviet Union. Maltese, on the island of Malta, has developed into a separate language, spoken by about 300,000. A typical feature of the group is a verbal system that is very similar to that of the Northern Central group (with minor differences) but that developed tenses instead of aspects from the late Middle Stage (*e.g.*, in the Egyptian dialect, from prepositional constructions). There were three cases of the noun, but declension was lost at the late Middle and New Stage. Also characteristic of South Central Semitic are the article *al-* and a strongly developed system of internal inflection with the plural mostly of the *pluralis fractus* type ("broken plural," in which plurality is shown by means of internal vowel changes). The Proto-Semitic phonological system has been on the whole well preserved, but **ś* has become *s*, **ṭh* has become *ẓ*, and other similar changes.

Southern Peripheral Semitic. To the Southern Peripheral group of the Ancient to Middle Stage belong the South Arabian dialects, Sabaean (*cf.* Sheba), Minaean, Qatabānian, and Ḥaḍramawtian, spoken from the 1st millennium BC to the 1st millennium AD. The Middle Stage is represented by Geʿez (Geez or Gəʿəz), or Ethiopic, found in northern Ethiopia in the 1st millennium AD; and the New Stage by the South Arabian group, including Mahrī, Shaḥrī (Eḥkalī), Ḥarsūsī, and Baṭḥarī on the Arabian shore of the Indian Ocean, and Suquṭrī (possibly a dialect of Mahrī) on the island of Socotra, with the total number of speakers probably being around 50,000. Also of the New Stage is the Ethiopic group, consisting of three subgroups: North Ethiopic, Central Ethiopic, and the Gurage subgroup. North Ethiopic, nearest to Geʿez, includes Tigrinya (Tigrai) and Tigre, spoken in northern Ethiopia and Eritrea by 4,500,000 speakers. Within the Central Ethiopic group (10,000,000 speakers) are Amharic, the official language of Ethiopia, the near-extinct Argobba language, and the entirely extinct Gafat. The Gurage cluster of languages is found south and east of Addis Ababa and has 650,000 speakers. In all, there are somewhat more than 15,000,000 speakers of Southern Peripheral Semitic languages.

Ethiopic group of languages

Typical features of the group include traces of two types of verbal forms with prefixed actor marker (one type with full vowels and the other with reduced vowel schemes). In other respects the verbal system is as in South and North Central Semitic; considerable innovations, however, have developed at the New Stage, especially in the Ethiopic group (with a Cushitic substratum). Declension was lost from the Middle Stage. Phonetic development is as in Arabic, but more of the ancient consonants were lost. The Ethiopic group has lost most of the pharyngeal and laryngeal consonants.

Table 49: The Arabic Alphabet and Numerals

consonants					equivalents		approximate pronunciation, classical Arabic
alone	initial	medial	final	name	EB preferred	alternatives	
ا	ا	ـا	ـا	alif	*		*
ب	بـ	ـبـ	ـب	bāʾ	b		*b*a*b*y
ت	تـ	ـتـ	ـت	tāʾ	t		*t*ie†
ث	ثـ	ـثـ	ـث	thāʾ	th	t͟h	*th*in
ج	جـ	ـجـ	ـج	jīm	j	d͟j	*j*ob
ح	حـ	ـحـ	ـح	ḥāʾ	h	ḥ	‡
خ	خـ	ـخـ	ـخ	khāʾ	kh	k͟h	Ger. Bu*ch*§
د	د	ـد	ـد	dāl	d		*d*i*d*†
ذ	ذ	ـذ	ـذ	dhāl	dh	d͟h	*th*en
ر	ر	ـر	ـر	rāʾ	r		e*rr*or (trilled)
ز	ز	ـز	ـز	zāʾ	z		*z*one
س	سـ	ـسـ	ـس	sīn	s		*s*and
ش	شـ	ـشـ	ـش	shīn	sh	s͟h, š	*sh*y
ص	صـ	ـصـ	ـص	ṣād	ṣ	ş	‖
ض	ضـ	ـضـ	ـض	ḍād	ḍ	ḑ	
ط	طـ	ـطـ	ـط	ṭāʾ	ṭ	ţ	
ظ	ظـ	ـظـ	ـظ	ẓāʾ	ẓ	z̧	
ع	عـ	ـعـ	ـع	ʿayn	ʿ		¶
غ	غـ	ـغـ	ـغ	ghayn	gh	g͟h	Fr. *r*ien
ف	فـ	ـفـ	ـف	fāʾ	f		*f*i*f*ty
ق	قـ	ـقـ	ـق	qāf	q	ķ	♀
ك	كـ	ـكـ	ـك	kāf	k	k	*k*in
ل	لـ	ـلـ	ـل	lām	l		*l*i*l*y†
م	مـ	ـمـ	ـم	mīm	m		*m*ai*m*
ن	نـ	ـنـ	ـن	nūn	n		*n*o*t*
ه	هـ	ـهـ	ـه	hāʾ	h		*h*at
و	و	ـو	ـو	wāw	w		*w*atch ő
ي	يـ	ـيـ	ـي	yāʾ	y		*y*et ☐
ء	•			hamzah	initial, omit; medial and final, ʾ		*

vowels, diphthongs, and special diacritical marks		equivalents		approximate pronunciation, classical Arabic
letter	name	EB preferred	alternatives	
ـَ	fatḥah	a	e	*a*t
ـُ	ḍammah	u		f*oo*t
ـِ	kasrah	i		*i*f
ـَا	long fatḥah (alif)	ā		*a*dd, f*a*ther ◇
ـُو	long ḍammah (wāw)	ū		f*oo*d
ـِي	long kasrah (yāʾ)	ī		*e*ve
ـَوْ	fatḥah wāw sukūn	aw		*ou*t
ـَيْ	fatḥah yāʾ sukūn	ay	ai, ei	*i*ce
ى	alif maqṣūrah	ā	á	*a*dd, f*a*ther ◇
ة	tāʾ marbūṭah	-ah or -at	-a or -at	▲
ٱ	hamzat al-waṣl	restore a	ʾ	✚
آ	alif maddah	ā	a	*a*dd, f*a*ther ◇
ـِيّ , يّ	yāʾ shaddah followed by short vowel	īy-	iyy-	*e*ve⊕
ـُوّ	wāw shaddah followed by short vowel	ūw-	uww-	f*oo*d⊕

numerals

Arabic	westernized Arabic	Arabic	westernized Arabic	Arabic	westernized Arabic
١	1	١٢	12	٢٣	23
٢	2	١٣	13	٢٤	24
٣	3	١٤	14	٢٥	25
٤	4	١٥	15	٢٦	26
٥	5	١٦	16	٢٧	27
٦	6	١٧	17	٢٨	28
٧	7	١٨	18	٢٩	29
٨	8	١٩	19	٣٠	30
٩	9	٢٠	20	١٠٠	100
١٠	10	٢١	21	١٠٠٠	1,000
١١	11	٢٢	22		

**Alif* has no consonantal value of its own; in transliteration, omit at the beginning of a word. *Hamzah* is pronounced as a glottal stop, as in Cockney or New York "bottle." As a vowel, *alif* is pronounced *a*dd. †Pronounced dentally. ‡A pharyngeal fricative. §A velar fricative. ‖It is impossible by English examples to indicate the difference from pronunciation of *ş*, *ḑ*, *ţ*, and *z̧*. The back of the tongue is raised and the pharynx is constricted (velarization), in addition to the regular articulation of these consonants. ¶A contraction of the throat (a pharyngealized vowel that is considered a consonant in Arabic). ♀A uvular stop; a *k* sound produced farther back in the throat than any English *k*. őAs a consonant, *w*atch; as a vowel, f*oo*d. ☐As a consonant, *y*et; as a vowel, *e*ve. ◇ Pronunciation varies from place to place and also depending on the accompanying consonant. ▲Pronounced as *t* in certain grammatical constructions; otherwise pronounced as silent *h*. ✚Initial vowel elision; that is, in pronunciation the *a* is simply omitted and the preceding vowel is pronounced with the following consonant. ⊕A long vowel sound followed by appropriate consonant sound. See also *footnotes* ő *and* ☐.

Table 50: The Hebrew Alphabet

consonants			numerical value†	transliteration		approximate Israeli Sefardic pronunciation
printed*	written*	name		EB preferred	alternatives	
א	א	alef	1	ʼ; omit at beginning of word		glottal stop or silent
בּ	בּ	bet	2	b		*boy*
ב	ב	vet		v	bh, b̲	*vend*
ג	ג	gimel	3	g		*girl*
ד	ד	dalet	4	d		*dove*
ה	ה	he	5	h; omit at end of word unless written with dot		*how*; silent at end of word
ו	ו	waw (vav)	6	w‡	v	*vend*
ז	ז	zayin	7	z		*zebra*
ח	ח	ḥet	8	ḥ	h̲, ḥ, ch	Ger. Bu*ch*
ט	ט	ṭet	9	ṭ	t	*toy*
י	י	yod	10	y§		*yet*
כּ	כּ	kaf	20	k		*key*
כ (ך)	כ (ך)	khaf		kh		Ger. Bu*ch*
ל	ל	lamed	30	l		*leg*
מ (ם)	מ (ם)	mem	40	m		*member*
נ (ן)	נ (ן)	nun	50	n		*now*
ס	ס	samekh	60	s		*so*
ע	ע	ʻayin	70	ʻ		glottal stop‖
פּ	פּ	pe	80	p		*paper*
פ (ף)	פ (ף)	fe		f	ph	*fan*
צ (ץ)	צ (ץ)	tzade	90	tz	z̲, z, ẓ, ṣ, ts	pe*ts*
ק	ק	qof	100	q	k, ḳ	*key*
ר	ר	resh	200	r		Fr. *r*ien: or trilled *r*
שׁ	שׁ	shin	300	sh	š	*shoe*
שׂ	שׂ	sin		s	ś	*so*
תּ	תּ	taw	400	t		*toy*
ת	ת			t	th	*toy* (Ashkenazi *so*)

vowels		transliteration		approximate Israeli Sefardic pronunciation
sign¶	name	EB preferred	alternatives	
◌ַ	pataḥ	a		*father*
◌ֲ	ḥatef pataḥ	a	ă	*father*
◌ָ	qametz gadol ♀	a	ā	*father*
◌ֶ	segol	e	ä	*pet*
◌ֱ	ḥatef segol	e	ĕ	*pet*
◌ֵ	tzere ḥaser	e	ē	*pet*
◌ֵי	tzere male	e	ê	*they*
◌ְ	shewa naʻ	e	ᵉ(above line)	(silent)

vowels		transliteration		approximate Israeli Sefardic pronunciation
sign	name	EB preferred	alternatives	
◌ִ	ḥireq ḥaser	i		*feet*
◌ִי	ḥireq male	i	î	*feet*
◌ֳ	ḥaṭef qametz	o	ŏ	*cord*
◌ָ	qametz qaṭan	o		*cord*
◌ֹ	ḥolem ḥaser	o	ō	*cord*
וֹ	ḥolem male	o	ô	*cord*
וּ	shureq	u	û	*soon*
◌ֻ	qibbutz	u		*soon*

*Final forms in parentheses. †Hebrew has a "ciphered" numeral system in which the letters of the alphabet have numerical value and are used as numbers. For example, numbers 11 through 19 are written with the letter *yod* (10) plus the letter for 1, 2, etc. (except 15 which is written *ṭet* [9] plus *waw* [6], and 16 which is written *ṭet* [9] plus *zayin* [7]); the 20s are combinations of *kaf* (20) plus the letter for 1, 2, etc; 500 is written *taw* (400) plus *qof* (100). ‡Functions as both a consonant and a vowel. See *ḥolem male* and *shureq* in "vowel signs" table above. §Functions as both a consonant and a vowel. See *tzere male* and *ḥiriq male* in "vowel signs" table above. ‖In formal pronunciation, pronounced as a pharyngeal voiced fricative sound. ¶The long horizontal line represents the consonant; vowel signs are placed above, below, or to the left of it, as shown. ♀In Ashkenazi pronounced as in "ball." *Note:* In transliteration, a letter is usually doubled in medial position to indicate a strong *dagesh* (*dagesh* is indicated in Hebrew by a point in the middle of a letter). *Pataḥ* under final *he*, *ḥet*, and *ʻayin* is pronounced before the consonant under which it is written.

Historical and cultural background. Glottochronological methods, which attempt to measure degrees of differences between related languages by comparing a list of basic vocabulary items, indicate that the first group to separate from the Common Semitic ancestral language was Akkadian (Northern Peripheral group, *c.* 3300 BC) and the second was the Southern Peripheral group (second half of 3rd millennium BC). The Northern Central group had contacts for a long time with the Southern Central languages, and linguistic division within the North Central group is dated at the beginning or middle of the 2nd millennium BC. The relative position of Arabic to the other Semitic languages is not quite clear, probably because of its uninterrupted contacts with Aramaic and other nomadic Semitic groups for many centuries.

Semitic writing forms

The oldest of the attested Semitic languages, Akkadian, was the vehicle of a great ancient literature written in a logosyllabic cuneiform writing system of Sumerian origin. Records of other ancient Semitic languages exist in various forms. Amorite, another ancient Semitic language, is known from proper names; Ugaritic was written in a quasi-alphabetic cuneiform script unconnected with the Akkadian. The Canaanites of Phoenicia used a still undeciphered syllabic script, the Proto-Byblian, in the 2nd millennium BC, while those of Palestine and the Sinai Peninsula employed another undeciphered writing, the Sinaitic script, which may be alphabetic in nature. All the other Semites used and, for the most part, still use consonantal quasi-alphabets with no means or only imperfect means to distinguish the vowels. All such alphabets—of which the more important are the Hebrew, the Syriac, and the Arabic—are descended from the Phoenician linear quasi-alphabet of 22 signs, first attested at Byblos and externally similar to the Proto-Byblian script. (All the European alphabets are descendants of the Phoenician, and all the Asiatic alphabets are descendants of the Aramaic variants of the Phoenician.) From a South Arabian variant of the earliest Semitic alphabet the Ethiopians developed a syllabic writing still in use for the languages of Ethiopia. Maltese uses the Latin alphabet.

Two of the later Semitic languages, Hebrew and Arabic, have been the languages of great religions, Judaism and Islām. The religious significance of Hebrew explains why the language, although already partly replaced by an Aramaic vernacular in the everyday life of Palestine in the late 1st millennium BC and early 1st millennium AD, was still preserved as a literary language by the Jews who were expelled from Palestine by the Hellenistic kings and the Romans between the 3rd century BC and the 2nd century AD. It has been revived in a modernized form as a spoken and written language in Israel. Classical Arabic has been preserved as a literary language since the Arabic conquest of North Africa and the Near and Middle East in the 7th and 8th centuries AD. The language was used for literary purposes by Muslims of different nations all through the Middle Ages and is still used as a language of the school and the administration and as the spoken language of the educated in all Arabic countries, although the vernacular New Stage Arabic idioms are to a great extent mutually unintelligible. There is a great amount of literature—scholarly, religious, scientific, and fiction—both in Hebrew and in Arabic. Of the other Semitic languages, Syriac was and is the language of certain Eastern Christian sects and was the means by which the Greek tradition was passed on to the Arabs. Another Christian sect, that of the Monophysites of Ethiopia, used Geʿez (Ethiopic) and still retains it in ecclesiastical use, but the literary and other secular remains are less important.

Common Semitic sounds

Linguistic characteristics. The Semitic branch of this language family is characterized by several general features.

Phonology. In phonology, the emphatic stop consonants *ṭ* and *q* (from **ḳ*) were retained but not **ṗ*. The affricates of the parent language (which are begun as stops and released as fricative sounds), if they ever existed, were lost or replaced by sibilant and interdental sounds (which are symbolized as *s, ṣ, z; th, ṯh, dh*); the lateral sounds and the interdentals were subsequently lost in most languages. The labialized velar sounds (except in the Ethiopic group) and all postvelar stops were lost. As for the glottal, pharyngeal, and laryngeal consonants, six of them (*gh, kh, ʿ, ḥ, ʾ, h*) are retained in Arabic and were retained in the other Semitic languages at the Ancient Stage. Hebrew and Aramaic retain *ʿ, ḥ, ʾ*, and *h* (but only *kh*, and *h* in Modern Hebrew and in most New Aramaic dialects); later Ethiopic and Punic retained only *ʾ* and *h*, and Akkadian only *kh*, and *ʾ* (but *a* became *e* in words that formerly included *gh, ʿ*, or *ḥ*). The original six-vowel system changed everywhere as early as the Middle Stage; Arabic preserved it the longest.

Word formation. Word formation is achieved by an intricate system of vowel infixation, sometimes accompanied by a few suffixes or prefixes. Each pattern of infixation ("scheme"), in combination with a consonantal root, plus the affixes, has its own peculiar type of meanings. The Arabic noun *ma-KTaB-*, for example, means "place of writing, school," and *KaTTāB-* means "writer, scribe"; *KāTiB-*, a participle, means "writer, [the one] writing"; *ya-KTuB-u*, the imperfective form, is "he writes"; *yu-KaTTiB-u*, another imperfective, is "he writes, he teaches to write"; and *KaTaBa*, the perfective, means "he wrote." (The capital letters indicate the sounds of the consonantal root.) The need to correlate these diverse patterns with the basic meaning of the root resulted in the absence of important positional changes in the Semitic consonant sounds as well as in the comparative scarcity of borrowed terms, especially of verbs. The primary nouns, those not derived from verbal forms, are not included in this system of patterns, except, by analogy, in Arabic and the other Southern languages.

Morphology. In regard to morphology, the masculine gender marker is zero (*i.e.*, it has no structural marker), but traces of a *-u* can be detected; the feminine gender marker is *-a* or *-ā* or more usually *-(a-)t-*, although *-t-* belonged originally to another series of gender markers (in which there were more than two genders). The declension of the noun and pronoun was retained in the Ancient Stage of Semitic, with nominative, genitive, accusative, dative-locative, and locative-adverbial cases. The dative-locative ending was lost in Arabic, and the locative-adverbial form appeared only in Akkadian. There are traces of an earlier suffixed definite article, *-m(a)* or *-n(a)*, retained in Arabic as the marker of the indefinite form of the noun. Later new definite articles developed. The dual number was expressed at the Ancient Stage but became lost in the later stages. The plural of the noun is formed in North Semitic by lengthening the singular form. This means of expressing the plural exists also in South Semitic, but here it is to a great extent replaced by the so-called *pluralis fractus* "broken plural" (*e.g.*, Arabic *kalb-* "dog," *kilāb-* "dogs").

Dual and plural forms

The West Semitic languages (except, in part, for Ethiopic and the South Arabian dialects) have lost the old imperfective form with a full vowel scheme and have replaced it by the form next in frequency, namely, the subjunctive mood (*ya-qtul-u*) with a reduced vowel scheme. The stative verb form, still preserved in Akkadian, developed into a new perfective (Arabic *qatala* "he killed," *mariḍa* "he was ill"), leaving the form **ya-qtul*, which was originally the perfective and jussive (a form expressing a wish or command), for the jussive only.

Syntax. Typical of the Semitic languages are attributive constructions: (1) a construction in which the governing noun appears in a shortened form before the governed noun in the genitive, as well as (2) a construction in which the two nouns, each in its complete case form, are connected by a pronoun (*e.g.*, Old Akkadian *thu*, Aramaic *dhī*).

Vocabulary. As mentioned above, the system of word formation in Semitic does not favour borrowings, especially verbal ones. There are, however, a number of nouns borrowed from Sumerian in Akkadian; from Akkadian, Iranian, and Greek in Aramaic; from Persian and Turkic in Arabic; and from the Agau and other Cushitic languages in the Semitic languages of Ethiopia.

EGYPTIAN

The Egyptian branch of the Hamito-Semitic family includes only one language (with local dialects), namely, Egyptian. It can be differentiated into several stages. The

Ancient Stage, Old Egyptian, extended from before 3000 to *c.* 2200 BC; the transition from the Ancient to the Middle Stage, Middle Egyptian, lasted from *c.* 2200 to *c.* 1600 BC and, as a dead literary language, until *c.* 500 BC. The Middle Stage, Late Egyptian (also called Neo-Egyptian), is dated from *c.* 1550 to after *c.* 700 BC and the Demotic language between *c.* 700 BC and some time after AD 400. Finally, the New Stage, called Coptic, began in about the 2nd century AD and lasted at least to the 17th century and possibly, in some villages, until the 19th century. Thus, five literary dialects are differentiated. All these language periods refer to the written language, which often differed greatly from the spoken dialects. Coptic is still in ecclesiastical use (along with Arabic) among the Arabic-speaking Monophysite Christians of Egypt.

Character-istic Egyptian sound changes

Phonology. The Egyptian hieroglyphic writing was not adapted for expressing vowels. By the Coptic period, when an alphabetic writing came into use, the Egyptian vowel system had undergone so radical a change that the original vowels can be reconstructed only very approximately. In the consonantal system the loss of the emphatics (except *p* from **ṗ* and *q* from **ḳ*) is characteristic, as are the changes of **-r* (at end of syllable) to -ʾ, **li-* and **lu-* to *ỉ-*, **ki-* and **ku-* to *ṯ* (pronounced as *tch*), **gi-* and **gu-* to *ḏ* (pronounced *dj*). In some cases *ṯ* and *ḏ* apparently reflect the affricates of the parent language. In addition, the original lateral sounds were lost as well as the postvelar stops and labialized velars, and the system of spirants was simplified. Beginning with Middle Egyptian, *d, ḏ,* and *ṯ* developed gradually to *t,* and many final consonants (*e.g.,* *-t, -r*) were dropped.

Word formation, morphology, and syntax. Word formation in Egyptian was similar to the Semitic type, although probably less consistent. As for the inflection, there may have been only two cases of the noun in Old Egyptian. The actor case coincided with the genitive, and this may have been responsible for a drastic rearrangement of the entire verbal system. Of the original Hamito-Semitic verbal forms only the stative ("pseudo-participle") is preserved; its subject, when a pronoun, is in the ancient direct (zero) case. Verbal forms expressing action were replaced by attributive and prepositional constructions, with the person of the actor being expressed by a suffixed possessive pronoun or by a noun in the genitive. Stem modifications are less developed in Egyptian than in Semitic; a habitative form with reduplication (repetition) of the third consonant of the root exists along with the normal imperfective of the main stem. In the later periods a new complicated system of secondary verbal tenses developed. Masculine gender was marked by zero (the absence of any ending) or **-aw-*, feminine gender by *-at-*, plural masculine by **-ā-w-*, and plural feminine probably by **-ā-w-āt-*.

Typical of Egyptian syntax are a construction in which two nouns, each in its complete case form, are connected by a pronoun, called the *nota genitivi;* and a *status constructus,* in which the governing noun appears in a shortened form before the governed noun in the genitive.

Hiero-glyphic, hieratic, and demotic writing

Writing. The ancient Egyptian writing was a logosyllabic one, having symbols representing either complete words or syllables of words; identical signs were used for syllables with identical consonants but different vowels. According to the external form of the signs, the writing is classified as hieroglyphic when it is found on inscriptions on stone, metal, and other hard surfaces and as hieratic and later demotic when it is used for cursive writing on papyrus manuscripts. Typologically the three forms of writing are identical. Coptic was written in an alphabet based on Greek and partly on Demotic. There is a considerable literature in Egyptian and in Coptic (in the latter, mostly of a religious nature; see also WRITING: *Hieroglyphic writing*).

BERBER LANGUAGES

The Berber (Berbero-Libyan) branch is represented by a multitude of New Stage Berber dialects distributed all over North Africa, from the Siwa Oasis in the Arab Republic of Egypt to Senegal (about 11,000,000 speakers). The more important dialect clusters are Tamashek (Tuareg), in the central Sahara and south of the Niger; Shawia and Kabyle (Zouaouah), both in Algeria; Rif and Tamazight, predominantly in Morocco; Shluh (Tashelḥayt or Shilha), in Morocco and Mauritania; and Zenaga, in Senegal. Little is known of ancient Libyan, also called Numidian. It is attested by inscriptions found in Tunisia, Algeria, and elsewhere, dating from the times of the Roman Empire and written in a native consonantal quasi-alphabetic script still surviving in a modified form among the Tuaregs of Sahara. Whether the extinct language of the Guanches in the Canary Islands and of the Iberians of Spain belonged to the Berber branch or even to Hamito-Semitic is doubtful.

Ancient Numidian

Phonology. In the phonologies of these languages the vowels **a, *i, *u* were lost or reduced to *ə*, and **ā, *ī, *ū* became *a, i, u;* **w* and **y* may appear both as consonants and as vowels, and the emphatics are represented by *ḍ, gh* (but in reduplication *ṭṭ, qq*), and *ẓ*. The system of spirants has been simplified but retains *š* (*sh*) and *ž* (*zh*) sounds. Interdentals, laterals, and affricates were lost.

Word formation and morphology. Except in the verb, there are only traces of the internal inflection type of word formation characteristic of the Semitic branch. Among grammatical features, a former article no longer retaining its determinative function (masculine **hā-*, plural **hī*, feminine **tā-*, plural **tī-*) is prefixed under certain conditions to the noun, displacing the prefixed markers of gender, *w-* and *t-*. These latter gender markers are at present used in a form of the noun as an attribute or as a subject of a verb when following the predicate in the sentence. The plural of the noun is masculine *-ən* and *-an* and feminine *-in.* A *pluralis fractus* "broken plural" has also developed (mostly an infixation of *-a-*). The perfective of the main verbal stem also has a habitative form (reduplication of the second consonant of the root, or prefixation of *-tt-* to the word base). Tamashek has several verbal tenses.

There is little or no intelligibility between the dialects, except for historically neighbouring ones. A great number of Arabic borrowings are evident in most dialects; there are also a number of borrowings from Punic, Latin, and from the languages south of Sahara.

CUSHITIC LANGUAGES

The Cushitic branch goes back to a reconstructed Common Cushitic parent language; this, according to the Soviet scholar A.B. Dolgopolsky, was the dialect of Common Hamito-Semitic that best preserved the original phonological system. Whether or not West Cushitic (Omotic) is a descendant of Common Cushitic is not clear. The Cushitic languages are all from the New Stage and have about 16,000,000 speakers.

The five Cushitic language groups

Languages of the group. The Cushitic languages, including the West Cushitic group, can be subdivided into five groups. The Northern group is represented by Beja, or Bedawiye, spoken mainly in The Sudan close to the Red Sea and also in Eritrea; it has about 1,300,000 speakers. Typical linguistic features include the scanty representation of affricates, velars changing partly to ʾ, and postvelar consonants changing to *h.* Two or, in some cases, three verbal forms with prefixed actor markers exist ("strong conjugation"), but many verbs are conjugated by suffixes (developed from an auxiliary verb with prefix conjugation). There are stem modifications similar to those of Berber in the strong conjugation, formed by suffixes in the other verbs (this is also typical of the other Cushitic languages). In addition, declension of the noun, with traceable relics of the ancient type similar to the Semitic, also can be seen.

The Eastern group has several subgroups. The highland languages, spoken east of Addis Ababa, include Hadya-Libide (900,000 speakers), Kambata (300,000 speakers), Sidamo (1,100,000 speakers), Darasa (300,000 speakers), Burji, and some related languages. The total number of speakers of this subgroup is about 2,600,000. The other subgroups include Saho-Afar in Eritrea, northeast Ethiopia, and Djibouti, with 750,000 or more speakers; the Somali subgroup, with Somali, Bayso, Rendile, and other languages in Somalia, eastern Ethiopia, and eastern Kenya, having a total of more than 5,000,000 speakers; the Gallinya subgroup, comprising Oromo (Gallinya, with several dialects) in western, central, and southern Ethiopia and northern and eastern Kenya; the subgroup of Arbore, Dathanaic (Geleba), and other languages, together having about 13,000,000 speakers; Konso, Gidole, and related dialects, with about 80,000 speakers in western Ethiopia; Warazi (Warize) and related languages also in western Ethiopia, with about 50,000 speakers; and Mogogodo of northern Kenya. Typical features of the group are the presence of emphatic affricate sounds and the change of postvelar sounds to ʼ and *h;* in some languages the older **l* sound is represented as *j* or *r,* and **r* as *r, d,* or *n.* The number of verbs of the "strong conjugation" is very small in some languages and nonexistent in others. In addition, there are grammatical genders differing from the ancient type.

The Central, or Agau, group includes languages or dialects dispersed over Ethiopia. They include Bilin, Khamta, Awngi, and Kemant (Qimant), among others, and are spoken by more than 100,000 people. The Quara dialect, spoken formerly by the Falashas, an Ethiopian Jewish ethnic group, is now extinct. Although the vocabulary of all Agau dialects is very similar, there is little mutual intelligibility as a result of the dissimilarity in the phonetic reflexes of the Proto-Cushitic sounds and the strong influence of Ethiopic and Amharic.

The Southern group, located in Tanzania, south of the Equator, includes Iraqw and its related dialects, Asa, Ngomwia, and others. Characteristic of the group is the loss, for the most part, of emphatic consonants. The laterals, however, are partly preserved, as are the pharyngeal ʻ and a few of the affricates. Both **l* and **r* are reflected as *l-* and *-r-.* In spite of numerous innovations as a result of substratum influence, there are many similarities with Eastern Cushitic in grammar.

The Omotic languages

The Western group, also called the Omotic branch by some scholars, encompasses Ometo, a dialect cluster including the Walamo language, with about 1,600,000 speakers; Janjero (Yamma), Bworo (Shinasha, Gonga), Anfillo (Southern Mao), Benesho-She (Gimira), Ari-Banna, and others, all of which are languages with small numbers of speakers, perhaps about 120,000 in all; and Kafa (Kaficho)-Mocha, with more than 200,000 speakers. All of these languages are spoken along the western border of Ethiopia and in northern Kenya. Typical features include the change of **s* to *š,* the preservation of most affricates, but the loss of laterals and of all postvelar, pharyngeal, and laryngeal fricatives. The sonants **l, *r,* and **n* are usually represented alike as *n-* and *-r-.* Some languages have tones that serve to differentiate word meaning. Also characteristic are drastic innovations in the pronoun and the verb. Traces of the genders are usually represented as masculine *-ō* (from **-aw*) and feminine *-ā* and *-ē* (perhaps from **-at* or **-ay*).

There have been some attempts to create a written language for Oromo and especially for Somali on the basis of Ethiopic, Latin, or Arabic writing. An original Somali writing system was invented in the beginning of the 20th century, but at present Somali is written in the Arabic alphabet.

Linguistic characteristics. *Phonology.* All the Hamito-Semitic groups of consonants were preserved in Common Cushitic, and separate reflexes of each group can be traced in the different Cushitic languages. Because of an imperfect development of the system of word formation by vowel infixation, however, the stability of the consonantal root was not as necessary for correlation of forms as in Semitic. The reflexes of the sounds of the protolanguage in the individual Cushitic languages therefore depend to a great extent on positional circumstances; thus, a Proto-Cushitic **c,* pronounced as *ts,* may have developed into a *d-* in an initial position and an *-s-* in an intervocalic or final position, and so forth. Emphatics are mostly preserved (*ḍ, ç* or *č̣, ḳ,* etc.); **p̣* is distinguished from **p* by different reflexes (Omotic partly retains *p̣*). Affricates (and also *d, š, s,* etc.) represent what in Semitic are interdental consonants.

Morphology. Verbal conjugation by means of prefixed actor markers is preserved only in a part of the verbs or else in traces; in most of the verbs it is replaced by a new system of conjugation (originally a combination of verbal noun plus prefix conjugation of an auxiliary verb). Two genders (masculine **-w,* feminine **-t*) and traces of noun declension can be observed; partial and sometimes complete reduplication of stems is used as a means of expressing the plural, along with the means known from the other branches. The pronominal system (except in Omotic) is very close to that of Semitic. In vocabulary, there are many borrowings from Ethiopic, Amharic, Arabic, and Nilo-Saharan.

CHADIC LANGUAGES

Hausa

Languages of the group. Of the Chadic (Chado-Hamitic) branch the most important language by far is Hausa. Its approximately 22,000,000 speakers live in northern Nigeria, in the Republic of the Niger, in the northern part of Ghana, in Cameroon, and in parts of Togo, Benin (Dahomey), the Chad Republic, and the Central African Republic. Hausa is also spoken as a second language by many speakers of other African languages. All the other languages of the Chadic branch are spoken by smaller ethnic groups; for example, the Bura, with about 1,500,000 people; the Mandara, with about 670,000; the Angas, with about 500,000; the Bolewa and Karekare, numbering about 220,000; and the Kanakuru, with about 150,000. In regard to the classification of the Chadic languages, some units are at times classified as languages but should preferably be treated as dialects, and vice versa, and the information on many languages (dialects?) is very scanty. Thus, all attempts at classification must be regarded as provisional only. The more important of the approximately 150 "languages" follow.

1. Western group: Hausa, Gwandara, Ngizim, Bedde (Bade), and related languages; Warjawa, Afawa (Paʼa), and related languages; Gezawa, Seiyawa, Barawa of the Dass region; Bolewa, Karekare, Kanakuru, and related languages; Angas, Sura, Ankwe, Gerka, and related languages; Maha (?).
2. Ron group: Fyer, Bokkos, Daffo-Butura, Sha, and Kulere.
3. Kotoko group: Logone, Buduma, Afade, Gulfei, and related languages and dialects.
4. Musgu (Musgum).
5. Masa group: Masa (Banana), Bana, Kulung; Mussoi (?); Marba (?); and Dari (?).
6. Eastern group: Somrai and related languages; Gaberi and related languages and dialects; Sokoro and related languages; Modgel; Tuburi (Kera); Mubi; Dangla-Jegu, Jonkor, and related languages.

The affiliation of the following languages and dialects to the Hamito-Semitic family has been questioned by some scholars:

7. Tera, Jera; Hona, Gaʼanda, and related languages; Bata, Gundu, and related languages; Margi, Bura, Chibak, and related languages; Higi (Hiji) and related languages; Laamang (Hidkala); Mandara (Wandala), Glavda, Yawotatakha, Sukur, and related languages.
8. Daba, Hina, Gauar, Musgoi (?); Matakam, Mofu, Gisiga, and related languages.
9. Gidder.

The total number of speakers of Chadic languages is probably about 26,000,000.

It is probable that the linguistic area of the Chadic branch formerly extended farther to the east, thus contacting the Omotic (Western Cushitic) group.

No Chadic language except Hausa has been reduced to writing; for Hausa, Arabic writing began to be used in the 16th century, and now a modification of the Latin alphabet is used.

Linguistic characteristics. In relation to the original phonological system of Proto-Hamito-Semitic, there are many missing sounds in the individual Chadic systems, but a comparison of the different Chadic groups shows that all the distinctive sounds of the parent language are represented in one way or another. Typical of all Chadic languages are tones serving to differentiate meaning in otherwise identical words.

The verb is for the most part a combination of an auxiliary verb (prefix conjugation) followed by a verbal noun (the reverse order is typical for Cushitic); the nominal part often has a vowel suffix pointing out certain qualities of the verb (transitiveness, intransitiveness, and so on). Verbal stem modifications exist, being expressed by such devices as reduplication and suffixes, but in most languages they are not strongly developed. Most of the languages have more or less clearly expressed genders but no declension. Plural is shown as in Cushitic and Berber. Chadic languages have innovations in the pronominal system. In vocabulary there are borrowings from English, Arabic (especially in Hausa), Fulani, and East Sudanic.

(I.M.D.)

KOREAN LANGUAGE

Korean is spoken by more than 72 million people, of whom 45 million live in South Korea and 24 million in North Korea. There are more than 2 million speakers in China, more than 700,000 in the United States, and somewhat fewer in Japan. More than 200,000 Koreans live in Kazakstan and Uzbekistan. Korean is the official language of both South Korea (Republic of Korea) and North Korea (Democratic People's Republic of Korea). The two Koreas differ in minor matters of spelling, alphabetization, and vocabulary choice (including the names of the letters), but both essentially endorse the unified standards proposed by the Korean Language Society in 1933.

LINGUISTIC HISTORY AND WRITING SYSTEMS

General considerations. While much is known about Middle Korean, the language spoken in the 15th century (when the script was invented), information about the language before that time is limited. Several hundred words of early Middle Korean were written with phonograms in the vocabularies compiled by the Chinese as far back as 1103. A still earlier form of the language, sometimes called Old Korean, has been inferred from place-names and from the 25 poems (called *hyangga*) that were composed as early as the 10th century and reflect the language of the Silla kingdom. Written with Chinese characters used in various ways to stand for Korean meanings and sounds, the poems are difficult to decipher, and there is no consensus on the interpretation of the content.

Nor is there general agreement on the relationship of Korean to other languages. The most likely relationships proposed are to Japanese and to the languages of the Altaic group: Turkic, Mongolian, and especially Tungus (-Manchu-Jurchen).

Writing and transcriptions. When Korean words are cited in English and other languages they are transcribed in a variety of ways, as can be seen from the spellings seen for a popular Korean surname: I, Yi, Lee, Li, Ree, Ri, Rhee, Rie, Ni, and so on. For English speakers the most popular transcription is that of the McCune-Reischauer system, which writes words more or less as they sound to the American ear. Despite its clumsiness, McCune-Reischauer is the system used in this description, and following that system the common surname is written Yi; it sounds like the English name of the letter *e*. In citing sentences, many linguists prefer the Yale romanization, which more accurately reflects the Korean orthography and avoids the need for diacritics to mark vowel distinctions. For a comparison of the two systems, see Table 51.

The writing system dates from 1443, and for many years it was known as *Ŏnmun* 'vernacular script,' though in South Korea it is now called Hangul (*han'gŭl;* or Hankul in the Yale romanization) and in North Korea Chosŏn kŭl(tcha), Chosŏn mun(tcha), or just Chosŏn mal 'Korean.' Very simple symbols are provided for each of the phonemes. Words can be spelled by putting these symbols one after another, as most writing systems do, but Koreans have preferred to group the symbols into square blocks like Chinese characters. The first element in the block is the initial consonant; if the syllable begins with a vowel, a small circle serves as a zero initial. What follows, either to the left or below (or both) is the vowel nucleus, which may be simple or complex (originally a diphthong or triphthong). An optional final element at the bottom (called *patch'im*) writes a final consonant or a cluster of two consonants. The 15th-century script had a few additional consonant letters that became obsolete in the following centuries and an additional vowel distinction that survived in the spellings until 1933; that vowel is usually transcribed as *ă.* On Cheju Island, where the distinction is maintained, the phoneme is pronounced [ɔ], very close to the modern Seoul version of the vowel transcribed *ŏ*, which in many parts of the country is still pronounced [ə]. That accounts for the first vowel of the usual spelling Seoul (= Sŏul), based on a French system of romanization, and for the use of the letter *e* to write *ŏ* in the Yale system.

The earlier language had a distinctive musical accent. In the far south and the northeast, the accent is still maintained as distinctions of pitch, vowel length, or a combination of the two. In the 15th century, low-pitched syllables were left unmarked but a dot was placed to the left of the high-pitched syllables and a double dot (like a colon) was put beside syllables that rose from low to high. The rising accent was maintained as vowel length in central Korea after the other distinctions eroded, but it, too, is vanishing in modern Seoul, even in initial syllables, where it has persisted longest. Like French, Seoul Korean no longer uses accent to distinguish words. The few apparent exceptions

Table 51: The Korean Alphabet (Hangul)

consonants	McCune-Reischauer romanization: initial	medial	final	Yale romanization: all positions	approximate equivalent sound in English
ㄱ	k	g	k	k	*k*en, a*g*ain, nec*k*
ㄲ	kk	kk	k	kk	s*k*in, as*k*ance, mas*k*
ㄴ	n	n	n	n	*n*o, a*n*y, o*n*
ㄷ	t	d	t	t	*t*o, la*d*y, ma*t*
ㄸ	tt	tt	—	tt	s*t*eam, as*t*ound, —
ㄹ	r	r*	l	l	*r*ed, eigh*t*y, pa*l*
ㅁ	m	m	m	m	*m*ow, a*m*ong, a*m*
ㅂ	p	b	p	p	*p*ie, a*b*ut, mo*p*
ㅃ	pp	pp	—	pp	s*p*y, ex*p*ose, —
ㅅ	s	s	t	s	*s*un, con*s*ole, cas*t*
ㅆ	ss	ss	t	ss	(it)'*s* OK, mi*ss*ay, mis*t*
ㅇ	(silent)	ng	ng	ng	—, si*ng*er, so*ng*
ㅈ	ch	j	t	c	*ch*ill, a*j*ar, ba*t*
ㅉ	tch	tch	—	cc	(it's) *j*am, it *j*umps, —
ㅊ	ch'	ch'	t	ch	(mu)*ch h*elp, su*ch h*ope, —
ㅋ	k'	k'	k	kh	(bac)*k h*ome, coo*kh*ouse, as*k*
ㅌ	t'	t'	t	th	(i)*t h*elps, a*t h*ome, a*t*
ㅍ	p'	p'	p	ph	(u)*p h*ere, u*ph*old, u*p*
ㅎ	h	h	h	—	*h*ope, a*h*ead, —

*When one syllable ends with [l] and the next begins with [r], the string is written ll and not lr.

vowels and diphthongs	McCune-Reischauer romanization	Yale romanization	approximate equivalent sound in English
ㅏ	a	a	*a*h, h*o*t
ㅐ	ae	ay	b*a*t (or b*e*t, b*ay*)
ㅒ	yae	yay	*ya*p (or *ye*t)
ㅓ	ŏ	e	s*aw* (or s*u*ng)
ㅔ	e	ey	s*e*t, h*ey*
ㅕ	yŏ	ye	*yaw*l (or *you*ng)
ㅖ	ye	yey	*ye*t, *Ya*le
ㅗ	o	o (or wo)	h*o*me, kn*ow*
ㅘ	wa	wa	*wo*k
ㅙ	wae	way	*wa*x, s*wa*m (or *we*t, *way*)
ㅚ	oe	oy (or woy)	*we*t, *way* (or French *eu*x)
ㅛ	yo	yo (or ywo)	*yo*ke
ㅜ	u	wu	t*oo*l
ㅝ	wŏ	we	*wa*n (or *wo*n)
ㅞ	we	wey	*we*t, *way*, h*ey*
ㅟ	wi	wi	ki*wi*, *we* (or French h*ui*t)
ㅠ	yu	yu (or ywu)	*yu*le, *you*
ㅡ	ŭ	u	t*oo*k
ㅢ	ŭi	uy	= ŭ + i or ŭ-/-i
ㅣ	i	i	ch*i*l*i*, t*i*p, mar*i*ne

are due to intonation: *nu-ga wassŏ* (spoken with a rising pitch) 'Did someone come?', *nu-ga wassŏ* (spoken with a falling pitch) 'Who came?'.

Koreans began putting spaces between words in 1896. As in English, judgment varies on what constitutes a word rather than a phrase. Earlier, Koreans wrote syllables as distinct blocks but failed to separate words. That was the Chinese tradition, which is still alive in Japan, where the mixture of kanji (Chinese characters) and kana (syllabic symbols based on kanji) helps the eye detect phrase breaks. The Chinese comma and period (a hollow dot) are commonly used, and modern punctuation marks have been taken from English.

Korean borrowed many words from Classical Chinese, including most technical terms and about 10 percent of the basic nouns, such as *san* 'mountain' and *kang* 'river.' The borrowed words are sometimes written in Chinese characters, though that practice is increasingly avoided except when the characters are used as aids in explaining technical terms.

Korean spelling is complicated. Words are usually written morphophonemically rather than phonemically, so that a given element is seen in a constant form, even though its pronunciation may vary when it is joined with other elements. For example, the word for 'price' is always spelled *kaps* though it is pronounced /kap/ in isolation and /kam/ in *kaps-man* 'just the price.' See below *Linguistic characteristics: Assimilations*. From the 15th century on there has been a steady trend toward ignoring predictable alternants.

Digraphs and separators. All transcriptions of Korean include digraphs of one kind or another and use separators to distinguish a string of two letters in their separate values from their single value as a digraph. When no other mark (such as a hyphen or space) is in order, the McCune-Reischauer system uses the apostrophe to distinguish such pairs as *hangŏ* (= *hang-ŏ*) 'resistance' and *han'gŏ* (= *han-gŏ*, usually pronounced as if *hang-gŏ*) 'a cloistered life.'

LINGUISTIC CHARACTERISTICS

Vowels. The vowel nucleus consists of a simple vowel, which may be preceded by *y* or *w*. The McCune-Reischauer romanization puts a breve (˘) over the letters *u* and *o* to distinguish the originally unrounded vowels [ɨ] and [ə] (= Seoul [ɔ]) from their rounded counterparts [u] and [o]. (Unrounded vowels are said with a tight smile; rounded vowels with pursed lips.) The Yale romanization uses the letter *u* for the unrounded [ɨ] and writes [u] as *wu* but encourages the omission of the *w* after *p, ph* (= *p'*), *pp, m,* and *y,* where the rounding has become nondistinctive in modern Korean. The front vowels transcribed as *e* = [e], and *ae* = [ɛ] or [æ] were originally diphthongs, as shown by the complex Hangul symbols reflected in the *ey* and *ay* of the Yale romanization. The vowel *ae* is no longer kept distinct from *e* in southern Korea and the distinction is virtually lost in modern Seoul, though it is maintained in the spellings. Another old diphthong, originally [oy], is transcribed *oe* (Yale *oy*) and sometimes pronounced as a front rounded vowel [ö], though it commonly sounds the same as the less common diphthong *we*, which is often simplified to just *e*; the surname Ch'oe may be said as if spelled Ch'we or Ch'e. The old diphthong [uy] became modern *wi* [wi] and is so pronounced by most speakers, but some people use a front rounded vowel [ü]. The old diphthong *ŭi* [ɨy], into which *ăe* [əy] merged, was largely replaced by [ɨ] (initially) or [i], but it is maintained in spelling certain words of Chinese origin such as *ŭiŭi* [ɨ:i] 'meaning' and in writing the particle *-ŭi* [e] 'of' (but not the homonymous particle *-e* 'to' or 'in'). The older version as a diphthong has regained popularity in modern Seoul in words such as *ŭija* 'chair' (said as three syllables), probably as a result of "reading" pronunciations.

Consonants. When they are initial, the simple stops *p, t,* and *k* are pronounced much as in English (*p*ie, *t*ie, *k*ite), with light aspiration. When final, they are cut off with no release, as in one way of saying English "U*p*! Ou*t*! Bac*k*!". The affricate *ch* occurs in the word *chip* 'house,' which is pronounced with a sound intermediate between English *chip* and *cheap;* some speakers, especially before the back vowels, pronounce the affricate as a nonpalatalized [ts], and that is thought to have been its 15th-century pronunciation. Between voiced sounds (which include the vowels and *y, w, m, n, ng, l,* and *r*), the stops acquire voicing, and that feature is noted in the McCune-Reischauer romanization (but not in the Hangul spelling or the Yale romanization): *ip* 'mouth' but *ib-e* 'in the mouth,' *mat* 'firstborn' but *mad-adŭl* 'eldest son,' *ak* 'evil' but *ag-in* 'evil person.' Final *-p* sometimes represents a basic *p'* (*ap* 'front' but *ap'-e* 'in front') or *ps* [pss] (*kap* 'price' but *kaps-ŭl* 'the price [as object]'). A final *t* sometimes represents a basic *t'* (*mit* 'bottom' but *mit'-e* 'at the bottom'), *j* (*nat* 'daytime' but *naj-e* 'in the daytime'), or *ch'* (*kkot* 'flower' but *kkoch'-ida* 'it's a flower'); more often, however, a final *t* represents a basic *s* (*ot* 'garment' but *os-ŭl* 'the garment [as object]'). Some speakers regularize the basic forms of nouns (but not verb stems) so that for them the nonfinal *t* always represents basic *s;* they say *pas-ey* for the standard *pat'-ey* 'in the field.' The single liquid phoneme has two predictable pronunciations: a clear (and sometimes palatalized) lateral [l] when it is at the end of a syllable or doubled [-l:l-], otherwise (and also before *h*) the flap [r]. The McCune-Reischauer transcription writes *l* or *r* in accordance with the pronunciation, which is ignored by the Hangul spelling and the Yale romanization: *il* 'one' but *ir-wŏn* 'one Wŏn,' *nal* 'day' and *nal-lo* 'by the day' but *nar-e* 'on the day.' There are problems involving initial *r* and *n* that are reflected in newspaper references to "President Roh (pronounced No)." The Korean language borrowed some Chinese words beginning with a liquid ([l] in Chinese), and Koreans tried to pronounce them with an *r,* but they were generally successful only when the element was not initial in a word; when the liquid was initial they used an *n* instead. The second syllable of *toro* 'street' is the same element as the first syllable of *nosang* 'on the street' (which is spelled with an *n-* in South Korea but with an *r-* in North Korea). There is an added complication: in the south *ny-* and *ni-* (whether from original *n-* or from *r-*) dropped the nasal, and that is when the common surname borrowed from Chinese Lǐ became Yi. The name is still pronounced [ni] in parts of North Korea, though the reading pronunciation [ri] has spread in P'yŏngyang since 1945. In modern times, loanwords from English, Japanese, and Russian have brought in an initial [r], and that is usually pronounced as a flap.

Assimilations. The spoken syllables are fairly simple in structure. Each ends either in a vowel or in one of the voiced consonants *p, t, k, m, n, ng,* or *l*. When two syllables are put together, various changes may take place where they join. When a syllable that ends in a stop is followed by one that begins with a nasal, the stop assimilates: *chip* 'house' + *-man* 'only' sounds just like *chim* 'burden' + *-man* [čimman], and *kung-min* can mean either 'the people of the nation' (when the first syllable is *kuk-* 'nation') or 'the poor people' (when the first syllable is *kung-* 'poor'). Hangul spelling distinguishes such pairs by writing the basic forms. Before a velar (*k, k', kk*), the dental *n* is usually pronounced like the velar *ng* so that *kan'go* 'hardship' sounds like *kanggo* 'stable,' but that assimilation is ignored in both the spelling and the transcriptions. Both *-n* + *l-* and *-l* + *n-* are pronounced like *-l* + *l-*, so for the sound [-l:l-] one must know what is in the word to decide which of the three Hangul spellings to use.

Aspirated and reinforced consonants. English makes a two-way distinction of voiceless and voiced stops (*pip, bib; tat, dad; kick, gig*). In Korean, voicing is automatic, so that [p] and [b] form a single phoneme and are written with the same Hangul letter. Korean distinguishes two other kinds of obstruents (stops, or fricatives): heavily aspirated *p', t', k',* and *ch'* and reinforced (tense) *pp, tt, kk,* and *tch.* The standard language also has a tense sibilant *ss* in contrast with the lax (and somewhat aspirated) *s,* but many speakers maintain this distinction only at the beginning of a word or ignore it entirely, despite the spelling. Both kinds of *s* are palatalized before *i* or *y,* and the lax *s* sounds like English *sh,* so that the Silla kingdom is sometimes referred to as the Shilla kingdom. (The Yale transcription for this name, Sinla, shows the Hangul spelling.) The reinforced consonants, now written as geminates (duplicate letters), probably became distinctive through the reduction of clusters, such as Middle Korean *st, pst,* and *pt,* and in many words the heavily aspirated consonants seem to go back to earlier clusters with *h* or *k*. The clusters, in turn, were reduced from disyllabic strings by syncope (omitting the vowel). The simple aspirate *h* is often murmured or dropped between voiced sounds: *si(h)ŏm* 'test,' *annyŏng (h)ase-yo* 'How are you?' That accounts for the [r] in words like *sir(h)ŏm* 'experiment.' Before *i* the velar nasal is often reduced to no more than nasality: *annyŏng (h)i kase-yo* 'Good-bye to you who are leaving' is usually pronounced [annyɔ̃ĩgas].

Reinforcement. When *p, t, k, ch,* or *s* is preceded by a stop, it is automatically pronounced as reinforced (tense), but that sound feature is ignored both in the Hangul spellings and in the transcriptions: *ip-to* 'the mouth too' = /iptto/, *ot-kwa* 'with the garment' = /o(t)kkwa/, *kuk-poda* 'than the soup' = /kukppoda/, *hakcha* 'scholar' = /hak-tcha/, *iksal* 'joke' = /ikssal/. After the adnominal ending *-(ŭ)l* there is reinforcement unless a pause is inserted: *pol kot* 'the place to look' sounds just like *pol kkot* 'the flower to look at,' but they are spelled and transcribed differently. Other cases of reinforcement are less predictable and are variously treated or neglected in spellings and transcriptions. In some cases the reinforcement goes back to the Middle Korean particle *s* 'of.'

Liaison. When a Korean syllable that begins with a vowel is added to a syllable that ends in a consonant, that coda moves over to fill the empty onset slot of the second syllable: *chip* 'house' + *-e* 'to' is romanized *chib-e* but pronounced /či-be/, *pap* 'rice/meal' + *-ŭl* (direct object) is spelled *pab-ŭl* but said as /pa-bŭl/, and *tong-an* 'interval' is heard as /to-ŋan/.

Grammar. Korean sentences are very similar to those of Japanese, though the words sound quite different. Modifiers always precede what they modify. The unmarked order is subject + indirect object + direct object + predicate. Only the predicate is essential, and other information may be omitted. Actions are expressed by processive predicates (= verbs), such as *mŏgŏ* '[someone] eats [it]' and *anja* '[someone] sits,' characteristics by descriptive predicates (= adjectives), such as *tŏwŏ* '[it] is warm' and *cho(h)a* '[it] is good' (or 'I like it'). A special kind of descriptive, the closely attached copula (linking verb), predicates nouns, as in *ton–i(y)a* 'it's money' and *ch'a–'ya* 'it's tea.' The descriptive predicates can only make statements or questions, but the processive predicates can also make commands and suggestions, so that (depending on intonation) *anja* can mean 'I sit,' 'Will you sit?', 'Sit!', or 'Let's sit!'; *tŏwŏ* means only 'It's warm' or 'Is it warm?'. The predicate is an inflected form consisting of a stem + an ending: *mŏgŏ* consists of the stem *mŏk-* 'eat' and the infinitive ending *-ŏ,* which takes the shape *-a* when the preceding vowel is *a* or *o,* as in *anja* from *anch-* 'sit' and *cho(h)a* from *choh-* 'be good/liked,' and also (when the infinitive is sentence-final) in the copula *i(y)a* from the stem *i-*. A final particle *-yo* is often added to the infinitive to show friendly politeness, as in *mŏgŏ-yo* and *cho(h)a-yo;* the copula takes the shape *ie-yo* or *'(y)e-yo,* as in *ton–ie-yo* and *ch'a–'e-yo.* Nouns attach particles to show their roles in the sentence. The subject is marked by *-i* after a consonant or *-ga* after a vowel (but in the 15th century *-y* was used), and the direct object is marked by *-ŭl* after a consonant or *-lŭl* after a vowel. These case markers are often omitted or are masked by particles of focus such as *-do* and *-to* 'also, even,' used to highlight a word, and *-ŭn* (after a consonant) or *-nŭn* (after a vowel), used to background a topic. The indirect object is usually marked by *(h)ant'e* or (less casually) *ege,* as in *ŏmŏni-ga ttar-(h)ant'e ton-ŭl chue* 'The mother gives the daughter money.'

The negative copula expression converts [noun] + *i-* into [noun]-*i/-ga* + *ani–(i)-*, as in *ton-i ani–'ya* 'It isn't money' and *ch'a-ga ani–'ya* 'It isn't tea.' In short sentences verbs and adjectives can be negated by preposing the adverb *an(i),* as in *an mŏgŏ* 'I won't eat it,' but a more versatile device permits negation of even long sentences by converting the verb to an inflected form that ends in *-ji* or *-chi* (pronounced ch'i), followed by the auxiliary *anh-* 'not do/be,' a contraction from *ani ha-* (the negative adverb + the dummy verb *ha-* 'do/be'): *mŏkchi an(h)a* 'I don't eat it,' *chuji an(h)a* 'I don't give it,' *choch'i an(h)a* 'It's no good.' Predicates can be conjoined with endings meaning 'and' (*-go/-ko, -ŏ/-a–sŏ*) or 'but' (*-na/-ŭna, -ji/-chi-man*) and subordinated with endings that mean 'when/if' (such as *-myŏn/-ŭmyŏn*) or 'even if' (*-ŏ/a–do*).

The Korean language also has endings that adnominalize a predicate (make it modify a following noun), operating in the same manner as an English relative clause: *uri-(h)ant'e ton-ŭl chun saram* 'the people who gave us the money,' *saram-i uri-(h)ant'e chun ton* 'the money the people gave us,' *saram-i ton-ŭl chun uri* 'we to whom people gave money,' *saram-i uri-(h)ant'e ton-ŭl chun chib-i man(h)a* 'There are many houses where people gave us money.' Often the predicate marks the subject as someone special ("you" or "the teacher") by inserting the honorific marker *-(ŭ)si-,* reduced to *-(ŭ)sy-* before a vowel as in *chusyŏ-yo.* There are also ways to mark the predicate for tense and aspect.

The examples given here are all in an informal style, but there are several other styles. The formal style marks a statement with *-(sŭ)mnida,* a question with *-(sŭ)mnikka,* a command with *-(sŭp)sio,* and a suggestion with *-(sŭ)psida.* The forms in the plain/impersonal style are *-da* (or processive *-(nŭ)nda*) for statements, *-na* or *-ni* or *-(nŭ)n-ya/-ka* for questions, *-(ŭ)ra* or infinitive + *-ra* for commands, and *-ja/-cha* for suggestions. Instead of *-ŏ(-yo)* Koreans sometimes use *-ji(-yo)* to assert the speaker's involvement.

(S.E.M.)

JAPANESE LANGUAGE

With more than 125 million speakers, Japanese is one of the world's major languages, ranking ninth in terms of the number of speakers. It is primarily spoken throughout the Japanese archipelago; there are also some 1.5 million Japanese immigrants and their descendants living abroad, mainly in North and South America, who have varying degrees of proficiency in Japanese. Since the mid-20th century, no nation other than Japan has used Japanese as a first or a second language.

GENERAL CONSIDERATIONS

Hypotheses of genetic affiliation. Japanese is the only major language whose genetic affiliation is not known. The hypothesis relating Japanese to Korean remains the strongest, but other hypotheses also have been advanced. Some attempt to relate Japanese to the language groups of South Asia such as the Austronesian, the Austroasiatic, and the Tibeto-Burman family of the Sino-Tibetan languages. In the second half of the 20th century, efforts were focused more on the origins of the Japanese language than on its genetic affiliation per se; specifically, linguists attempted to reconcile some conflicting linguistic traits.

An increasingly popular theory along this line posits that the mixed nature of Japanese results from its Austronesian lexical substratum and the Altaic grammatical superstratum. According to one version of this hypothesis, a language of southern origin with a phonological system like those of Austronesian languages was spoken in Japan during the prehistoric Jōmon era (7500 BC to *c.* 250 BC). As the Yayoi culture was introduced to Japan from the Asiatic continent about 250 BC, a language of southern Korea began to spread eastward from the southern island of Kyushu along with this culture, which also introduced to Japan iron and bronze implements and the cultivation of rice. Because the migration from Korea did not take place on a large scale, the new language did not eradicate certain older lexical items, though it was able to change the grammatical structure of the existing language. Thus, this theory maintains, Japanese must be said to be genetically related to Korean (and perhaps ultimately to Altaic languages), though it contains Austronesian lexical residues.

Dialects. The country's geography, characterized by high mountain peaks and deep valleys as well as by small isolated islands, has fostered the development of various dialects throughout the archipelago. Different dialects are often mutually unintelligible; the speakers of the Kagoshima dialect of Kyushu are not understood by the majority of the people of the main island of Honshu. Likewise, northern dialect speakers from such places as Aomori and Akita are not understood by most people in metropolitan Tokyo or anywhere in western Japan. Japanese dialectologists agree that a major dialect boundary separates Okinawan dialects of the Ryukyu Islands from the rest of the mainland dialects. The latter are then divided into either three groups—Eastern, Western, and Kyushu dialects—or simply Eastern and Western dialects, the latter including the Kyushu group. Linguistic unification has been achieved by the spread of the *kyōtsū-go* "common language," which is based on the Tokyo dialect. A standardized written language has been a feature of compulsory education, which started in 1886. Modern mobility and mass media also have helped to level dialectal differences and have had a strong effect on the accelerated rate of the loss of local dialects.

Literary history. Written records of Japanese date to the 8th century, the oldest among them being the *Kojiki* (712; "Record of Ancient Matters"). If the history of the language were to be split in two, the division would fall somewhere between the 12th and 16th centuries, when the language shed most of its Old Japanese characteristics and acquired those of the modern language. It is common, however, to divide the 1,200-year history into four or five periods; Old Japanese (up to the 8th century), Late Old Japanese (9th–11th century), Middle Japanese (12th–16th century), Early Modern Japanese (17th–18th century), and Modern Japanese (19th century to the present).

Grammatical structure. Through the centuries, Japanese grammatical structure has remained remarkably stable, to the degree that with some basic training in the grammar of classical Japanese, modern readers can readily appreciate such classical literature as the *Man'yōshū* (compiled after 759; "Collection of Ten Thousand Leaves"), an anthology of Japanese verse; the *Tosa nikki* (935; *The Tosa Diary*); and the *Genji monogatari* (*c.* 1010; *The Tale of Genji*). Despite this stability, however, a number of features distinguish Old Japanese from Modern Japanese.

Phonology. Old Japanese is widely believed to have had eight vowels; in addition to the five vowels in modern use, /i, e, a, o, u/, the existence of three additional vowels /ï, ë, ö/ is assumed for Old Japanese. Some maintain, however, that Old Japanese had only five vowels and attribute the differences in vowel quality to the preceding consonants. There is also some indication that Old Japanese had a remnant form of vowel harmony. (Vowel harmony is said to exist when certain vowels call for other specific vowels within a certain domain, generally, within a word.) This possibility is stressed by the proponents of the theory that Japanese is related to the Altaic family, where vowel harmony is a widespread phenomenon. The wholesale shift of *p* to *h* (and to *w* between vowels) also took place relatively early, such that Modern Japanese has no native or Sino-Japanese word that begins with *p*. The remnant forms with the original *p* are seen among some Okinawan dialects; *e.g.,* Okinawan *pi* 'fire' and *pana* 'flower' correspond to the Tokyo forms *hi* and *hana*.

Syntax. Japanese syntax also has remained relatively stable, maintaining its characteristic subject–object–verb (SOV) sentence structure. A notable change in this domain is the obliteration of the distinction between the conclusive form—the finite form that concludes a sentence—and the noun-modifying form exhibited by certain predicates. For example, in early Japanese *otsu* and *tsuyoshi* were conclusive forms, respectively, of the verb 'to drop' and the adjective 'to be strong.' When these words were used as noun modifiers, the forms were inflected as *otsuru*, *tsuyoki*. This distinction between conclusive forms and noun-modifying forms played an important role in the phenomenon of syntactic concord that, for example, called for the noun-modifying forms of predicate even in concluding the predication when a subject or some other word was marked by particles such as the emphatic *zo* or the interrogative *ka* or *ya*. This system of syntactic concord deteriorated in Middle Japanese, and the distinction between the conclusive forms and the noun-modifying forms was also lost, the latter dominating the former. Such modern forms as *ochiru* 'to drop' and *tsuyoi* 'to be strong' are the descendants of the earlier noun-modifying forms.

A single most important development in the history of Japanese is the acquisition of the nativized writing systems that took place between the 8th and the 10th centuries. The Japanese vocabulary has been constantly enriched by loanwords—from Chinese in earlier times and from European languages in more recent history.

LINGUISTIC CHARACTERISTICS OF MODERN JAPANESE

Phonology. In Japanese phonology, two suprasegmental units—the syllable and the mora—must be recognized. A mora is a rhythmic unit based on length. It plays an important role especially in the accentual system, but its mundane utilization is most familiar in the composition of Japanese verse forms such as haiku and waka, in which lines are defined in terms of the number of moras; a haiku consists of three lines of five, seven, and five moras. A word such as *kantō* 'gallantly' consists of two syllables *kan* and *tō,* but a Japanese speaker further subdivides the word into the four units *ka*, *n*, *to*, and *o*, which correspond to the four letters of kana (see Table 52). In poetic compositions *kantō* is counted as having four, rather than two, rhythmic units and would be equivalent in length to a four-syllable, four-mora word such as *murasaki* 'purple.' While ordinary syllables include a vowel, moras need not. In addition to the moraic nasal seen in *kantō* above, there are several consonantal moras. These are the first of the double consonants—*e.g., kukkiri* 'distinctly,' *sappari* 'refreshing,' *katta* 'bought.' In the traditional phonemic analysis, the moraic nasal is analyzed as /N/ and the non-nasal moraic consonant as /Q/, and their phonetic values are determined by the following consonant (*e.g.,* /kaNpa/, pronounced kampa, 'cold wave,' /kaNtoo/, pronounced kantoo, 'gallantly,' /kaNkoo/, pronounced kaŋkoo, 'sightseeing,' /haQkiri/, pronounced hakkiri, 'clearly,' /yaQpari/, pronounced yappari, 'as expected'), except for an /N/ in final position, which is pronounced as a nasalized version of the preceding vowel (*e.g.,* /hoN/, pronounced hoõ, 'book,' /seN/, pronounced seẽ, 'thousand'). Long vowels count as two moras, and thus *ōkii* 'big' is a two-syllable (*ō-kii*), four-mora (*o-o-ki-i*) word.

The word-pitch accent system. Both moras and syllables play an important role in the Japanese accentual system, which can be characterized as a word-pitch accent system, in which each word (as contrasted with each syllable as in the prototypical tone languages of Southeast Asia) is associated with a distinct tone pattern. In Tokyo, for example, *hashi* with a high-low (HL) tone denotes 'chopstick,' but with a low-high (LH) tone it denotes 'bridge' or 'edge, end.' In Kyōto, on the other hand, *hashi* with a high-low tone means 'bridge,' and with a low-high tone it means 'chopstick,' whereas the word for 'edge, end' is pronounced with a flat high-high tone. The accentual system is one of the features that distinguishes one dialect from another, as each dialect has its own system, though certain dialects in the Tohoku region of northeastern Honshu and in Kyushu and some other areas show no pitch contrast.

In the majority of dialects, the pitch change occurs at the mora, not the syllable, boundary. The Tokyo form *kan* is a monosyllabic word, but, because it is dimoraic, pitch may change from high to low at the mora boundary, yielding *kan* (spoken with a high-low tone), which means 'official,' or (spoken with a low-high tone) 'sense.' Syllables, however, are units that determine the number of potential accentual distinctions, so that, given the possibility of unaccented forms, one-syllable words make two potential distinctions, two-syllable words three potential distinctions, and so forth. Thus, a monosyllabic word such as *e* can be either accented or unaccented and can be realized as a high-tone word (if accented) or as a low-tone word (if unaccented). The distinction, however, can be observed only when the form in question is followed by a particle such as the nominative particle *ga*; *e-ga* (LH) means 'handle [nominative]' and *e-ga* (HL) 'picture [nominative].' Since the number of potential distinctions is determined by the number of syllables in a word, monosyllabic and dimoraic words make only two potential distinctions. Thus, while there are accented *kan-ga* (high-low-low) 'official [nominative]' and unaccented *kan-ga* (low-high-high) 'sense [nominative],' there is no word pronounced with a low-high-low pitch. In other words, in the Tokyo dialect the number of potential accentual contrasts equals the number of syllables plus one. The absence of stress accent of the English type, the sequences of high-pitched moras as well as those of low-pitched moras, rather than alternating stressed and unstressed syllables, and the mora-timed characteristic together render Japanese speech rather monotonous compared to a stress-accent language like English or a true tone language like Chinese.

Phonemes. Japanese has the following phonemes: 5 vowels /i, e, a, o, u/, 16 consonants /p, t, k, b, d, g, s, h, z, r, m, n, w, j, N, Q/. The high back vowel *u* is unrounded [ɯ]. This and the other high vowel *i* tend to be devoiced between voiceless consonants or in final position after a voiceless consonant. The most pervasive phonological phenomena are palatalization and affrication, which turn *t*, *s*, *d/z*, and *h* into [tʃ], [ʃ], [dʒ], and [ç] before *i*, respectively, and *t* and *d/z* into [ts] and [dz] before *u*, respectively. The phoneme *h* also changes to [ɸ] before *u*. The effects of these processes are seen in inflected forms of verbs as well as in foreign loans—*e.g.*, /kat-e/ 'win [imperative]' /kat-anai/ 'win [negative],' /kat-oo/ 'win [cohortative],' /katʃ-imasɯ/ 'win [polite],' /kats-ɯ/ 'win [present]'; the English word *tool* becomes /tsɯɯrɯ/, *ticket* becomes /tʃiketto/, and *single* becomes /ʃiŋgɯrɯ/.

Grammatical structure. The first major part-of-speech division in Japanese falls between those elements that express concrete concepts (*e.g.*, nouns, verbs, adjectives) and those that express relational concepts (particles and suffixal auxiliary-like elements). The former elements may stand alone, constituting one-word sentences, whereas the latter always are attached to nouns and verbs and express grammatical concepts such as tense, the grammatical relations of subject and object, and the speaker's attitudes toward the proposition and toward the listener. Japanese verbs and adjectives conjugate and function as predicates without involving a copula (linking verb), whereas non-conjugating nouns and adjectival nominals (*e.g.*, *ganko* 'stubborn') require the copula *da* in their predication function—*e.g.*, *Tarō-ga ringo-o kau* (literally, Taro-[nominative] apple-[accusative] buy [present]) 'Taro buys an apple,' *Yama-ga taka-i* (literally, mountain-[nominative] high-[present]) 'The mountain is high,' *Tarō-wa sensei-da* (literally, Taro-[topic] teacher-[copula present]) 'Taro is a teacher,' *Tarō-wa ganko-da* (literally, Taro-[topic] stubborn-[copula present]) 'Taro is stubborn.' Predicates show no agreement for person, number, and gender. Nouns do not decline and do not indicate number or gender, while case distinctions are marked by enclitic particles (that is, particles attached to the end of the previous word), as in the examples above.

Japanese, as a consistent subject–object–verb (SOV) language, places modifiers before the modified, so that adjectives and relative clauses precede the modified nouns and adverbs come before verbs. A predicate complex consists of the stem followed by various suffixal elements expressing relational concepts. The order of these and other end-of-sentence, or sentence-final, elements reflect the ordering of meaning types from concrete to subjective to interpersonal; *e.g.*, *Ik-ase-rare-ru darō ka ne* (literally, go-[causative]-[passive]-[present], [conjecture], [question particle], [final particle]) 'Will (I) be made to go? What do you think?'

Elements recoverable from the context are freely omitted from Japanese, so that conversation abounds with sentence fragments, which may convey various meanings depending on the context—*e.g.*, *Kaita* (literally, write [past]) can mean 'I (he/she/they) wrote (a book, letters, etc.),' *Tarō to Jirō desu* (literally, Taro and Jiro [copula polite]) can mean 'Taro and Jiro came/played,' 'I met Taro and Jiro,' and so on. Some clues for recovering missing elements are provided for by means of honorific forms. When, for example, the verb *kaku* 'to write' is used in its subject honorific form—*kakareru* or *o-kaki-ni naru*—the writer referred to is not the speaker, but someone honoured as superior to the speaker. On the other hand, when the humble form *o-kaki suru* is used, the referent is likely to be the speaker. The addressee honorific form *kakimasu* is an index of the social relationship of the speaker to the listener, whereas the plain form *kaku* is used in addressing an equal, a social inferior, or an indefinite audience (as would be used, for example, in newspaper articles and books). The use of honorifics extends to the forms of personal address; one especially avoids use of *anata* 'you,' even in its honorific form *anata-sama,* when addressing a superior. The reference is usually omitted altogether, and the subject honorific form of the verb in combination with the addressee honorific form may simply be used, as in *O-iki-ni narimasu ka* 'Are [you] going?' If one must address a superior, that person's title or a kinship term is used, as in *Sensei-wa o-iki-ni narimasu ka* 'Are you going, Teacher?' Personal terms referring to the first person and particles that end the sentence also indicate the speaker's sex; opposed to the sex-neutral term *watakushi* for the first person are male forms *boku* and *ore* and typically female forms *watashi* and *atashi*. *Ze* and *zo* are final particles used by male speakers, while *wa* and *wa yo* are used exclusively by females.

The Japanese language exhibits a number of characteristic grammatical constructions not found in English and other European languages. An English sentence such as *John came* translates into two different expressions in Japanese. The sentence exhibiting the topic construction *John-wa kita* (John-[topic] came) contrasts with the basic sentence *John-ga kita* (John-[nominative] came), and the former is used when the referent of the *wa*-marked nominal (*i.e.*, John) is the topic of discussion, whereas the nontopic sentence simply describes the event in a neutral manner. The structure A-*wa* B-*da* (A-[topic] B-[copula]) bears a heavy functional load in Japanese. In addition to its basic identificational function (*e.g.*, *Kore-wa hon-da* 'This is a book'), the construction, supported by its context, is used to express a variety of meanings; *e.g.*, *Boku-wa ringo da* (literally, 'I am an apple') can mean 'I have decided to eat an apple,' 'I am going to pick apples,' and so on; *Boku-wa Kōbe-da* (literally, 'I am Kōbe') can mean 'I am going to Kōbe,' 'I am from Kōbe,' 'I am a fan of Kōbe,' 'I live in Kōbe,' 'I get off the train in Kōbe,' and so on.

Japanese also makes intransitive verbs passive; a passive sentence, such as *Boku-wa haha-ni shin-are-ta* (literally, I-[topic] mother-[dative] die-[passive-past]) 'I suffered from Mother's dying,' characteristically expresses the adversative meaning of suffering or inconvenience experienced by the subject referent (here, the male "I"). In addition to the regular passive of the type found in English, transitive verbs also produce adversative passive sentences—*e.g.*, *Boku-wa Hanako-ni piano-o hik-are-ta* (literally, I-[topic] Hanako-[dative] piano-[accusative] play-[passive-past]) 'I suffered from Hanako's piano-playing.'

Repetitive expressions abound in Japanese, and they profoundly affect both morphology and syntax. Examples of repetition include the use of syllable reduplication in various onomatopoeic expressions (*e.g.*, *ton-ton* symbolizes a light knocking sound, *don-don* symbolizes a heavy banging noise), the formation of plurals for certain nouns (*e.g.*, *yama-yama* 'mountains,' *hito-bito* 'people'), and the use of doubling in adverbial phrases for emphasis (*e.g.*, *hayaku-hayaku* 'quickly, quickly'). Additionally, the repetition of phrases yields a number of characteristic constructions of Japanese—*e.g.*, *yome-ba yomu-hodo omoshiroi* (literally, read-if read-to-the-extent interesting) 'the more (I) read, the more interesting it is,' *katta-ra katta-de ato-ga komaru* (literally, bought-if bought-at afterward-[nominative] suffer) 'If [I] bought [it] after all, then it would become troublesome afterward [I would regret it].'

Vocabulary. Japanese vocabulary consists of four lexical strata: native vocabulary, Sino-Japanese words, foreign loans, and onomatopoeic expressions. Each stratum is associated with phonological and semantic characteristics. The native vocabulary reflects the socioeconomic con-

Table 52: Japanese Kana*

simple *kana* symbols

No.	H	K	E	No.	H	K	E	No.	H	K	E	No.	H	K	E	No.	H	K	E
1.	あ	ア	a	2.	い	イ	i	3.	う	ウ	u	4.	え	エ	e	5.	お	オ	o
6.	か	カ	ka	7.	き	キ	ki	8.	く	ク	ku	9.	け	ケ	ke	10.	こ	コ	ko
11.	さ	サ	sa	12.	し	シ	shi	13.	す	ス	su	14.	せ	セ	se	15.	そ	ソ	so
16.	た	タ	ta	17.	ち	チ	chi	18.	つ	ツ	tsu†	19.	て	テ	te	20.	と	ト	to
21.	な	ナ	na	22.	に	ニ	ni	23.	ぬ	ヌ	nu	24.	ね	ネ	ne	25.	の	ノ	no
26.	は	ハ	ha	27.	ひ	ヒ	hi	28.	ふ	フ	fu	29.	へ	ヘ	he	30.	ほ	ホ	ho
31.	ま	マ	ma	32.	み	ミ	mi	33.	む	ム	mu	34.	め	メ	me	35.	も	モ	mo
36.	や	ヤ	ya					37.	ゆ	ユ	yu					38.	よ	ヨ	yo
39.	ら	ラ	ra	40.	り	リ	ri	41.	る	ル	ru	42.	れ	レ	re	43.	ろ	ロ	ro
44.	わ	ワ	wa													45.	を	ヲ	o
																46.	ん	ン	n(m)‡
47.	が	ガ	ga	48.	ぎ	ギ	gi	49.	ぐ	グ	gu	50.	げ	ゲ	ge	51.	ご	ゴ	go
52.	ざ	ザ	za	53.	じ	ジ	ji	54.	ず	ズ	zu	55.	ぜ	ゼ	ze	56.	ぞ	ゾ	zo
57.	だ	ダ	da	58.	ぢ	ヂ	ji	59.	づ	ヅ	zu	60.	で	デ	de	61.	ど	ド	do
62.	ば	バ	ba	63.	び	ビ	bi	64.	ぶ	ブ	bu	65.	べ	ベ	be	66.	ぼ	ボ	bo
67.	ぱ	パ	pa	68.	ぴ	ピ	pi	69.	ぷ	プ	pu	70.	ぺ	ペ	pe	71.	ぽ	ポ	po

digraphs representing single syllables

No.	H	K	E	No.	H	K	E	No.	H	K	E
1.	きゃ	キャ	kya	2.	きゅ	キュ	kyu	3.	きょ	キョ	kyo
4.	しゃ	シャ	sha	5.	しゅ	シュ	shu	6.	しょ	ショ	sho
7.	ちゃ	チャ	cha	8.	ちゅ	チュ	chu	9.	ちょ	チョ	cho
10.	にゃ	ニャ	nya	11.	にゅ	ニュ	nyu	12.	にょ	ニョ	nyo
13.	ひゃ	ヒャ	hya	14.	ひゅ	ヒュ	hyu	15.	ひょ	ヒョ	hyo
16.	みゃ	ミャ	mya	17.	みゅ	ミュ	myu	18.	みょ	ミョ	myo
19.	りゃ	リャ	rya	20.	りゅ	リュ	ryu	21.	りょ	リョ	ryo
22.	ぎゃ	ギャ	gya	23.	ぎゅ	ギュ	gyu	24.	ぎょ	ギョ	gyo
25.	じゃ	ジャ	ja	26.	じゅ	ジュ	ju	27.	じょ	ジョ	jo
28.	ぢゃ	ヂャ	ja	29.	ぢゅ	ヂュ	ju	30.	ぢょ	ヂョ	jo
31.	びゃ	ビャ	bya	32.	びゅ	ビュ	byu	33.	びょ	ビョ	byo
34.	ぴゃ	ピャ	pya	35.	ぴゅ	ピュ	pyu	36.	ぴょ	ピョ	pyo

hiragana trigraphs containing long vowels §

No.	H	E	No.	H	E
2.	きゅう	kyū	3.	きょう	kyō
5.	しゅう	shū	6.	しょう	shō
8.	ちゅう	chū	9.	ちょう	chō
11.	にゅう	nyū	12.	にょう	nyō
14.	ひゅう	hyū	15.	ひょう	hyō
17.	みゅう	myū	18.	みょう	myō
20.	りゅう	ryū	21.	りょう	ryō
23.	ぎゅう	gyū	24.	ぎょう	gyō
26.	じゅう	jū	27.	じょう	jō
29.	ぢゅう	jū	30.	ぢょう	jō
32.	びゅう	byū	33.	びょう	byō
35.	ぴゅう	pyū	36.	ぴょう	pyō

*H = *hiragana*; K = *katakana*; E = equivalent. Some *kana* undergo a change in pronunciation in specific situations.

†*Tsu* is also used to indicate a doubled consonant. In such cases the *tsu kana* is written slightly below the line (in horizontal texts) or slightly to the right of the line (in vertical texts) and sometimes also in slightly smaller script. Other *kana* are also positioned in this manner when they serve special functions.

‡Romanized *m* before *b*, *p*, and *m*.

§All five vowels have long forms, which in a Japanese text can be indicated in one of several ways. Romanized long vowels are indicated by macrons, except *i*, which is written *ii*.

cerns of traditional Japanese society, which were centred on farming and fishing. The words associated with rice, a staple food in Japan, clearly delineate the form or state of the rice to which they refer; the rice plant is *ine,* raw rice is *kome,* and cooked rice is either *gohan* or *meshi.* Both *gohan* and *meshi* are used to refer to meals in general, as an English speaker might use the word *bread* in the phrase 'our daily bread.' Another example of native vocabulary is the variety of names given to certain types of fish according to their size.

Some Chinese words are generally believed to have been introduced into Japan during the 1st century AD, or possibly before that. A systematic introduction of the Chinese language, however, occurred about AD 400, when Korean scholars brought Chinese books to Japan. Sino-Japanese words now constitute slightly more than 50 percent of the Japanese vocabulary, a proportion comparable to that of Latinate words in the English vocabulary. Both Chinese or Chinese-based words in Japanese and Latin or Latin-based words in English are also similar in their tendency to express abstract concepts and to make up a great part of the academic vocabulary. Contrary to what is suggested by the term *kan-go* 'Chinese word,' a large number of Sino-Japanese words were actually coined in Japan, using existing Chinese characters. Forms such as *shakai* 'society' and *kagaku* 'science' have been borrowed back into Chinese and adopted by Korean through the medium of shared Chinese characters.

Loanwords other than those constituting the stratum of Sino-Japanese words are lumped together as *gairai-go,* literally 'foreign-coming words.' In the contemporary Japanese vocabulary, English words dominate this category, with slightly more than 80 percent. Also evident are the linguistic legacies of 16th-century Portuguese, Spanish, and, in particular, Dutch missionaries and traders, as in such Modern Japanese words as *pan* 'bread' (from Portuguese *paõ*), *tabako* 'tobacco' (from Portuguese *tabaco*), *tenpura* '[English *tempura,* a deep-fried dish]' (from Portuguese *tempero*), *biiru* 'beer' (from Dutch *bier*), *penki* 'paint' (from Dutch *pek*), and *orugōru* 'music box' (from Dutch *orgel*). As illustrated in the last example, foreign loans are phonologically fully Japanized, with vowels appropriately inserted or appended and with occasional consonantal adjustments, although an initial *p*, which is lacking in Japanese, is left intact.

In fact, only the vocabularies of the native and the Sino-Japanese strata of Modern Japanese lack an initial *p*. It occurs quite frequently in the onomatopoeic vocabulary—*e.g., pachi-pachi* (referring to hand-clapping sounds), *piku-piki* (referring to a slight repetitive movement of an object), *piri-piri* (referring to a state of annoyance or irritation). As these examples suggest, Japanese sound symbolism encompasses not only mimetic expressions of natural sounds but also those that depict states, conditions, or manners of the external world as well as those symbolizing mental conditions or sensations. Sound-symbolic words permeate Japanese life, occurring in animated speech and abounding in literary works of all sorts.

Writing systems. The earliest attempts to write Japanese involved the use of not only Chinese characters but also Classical Chinese grammar, as is evident in the preface to the 8th-century *Kojiki.* Within some 50 years, by the time the *Man'yōshū* was completed, the Japanese had begun to use the sounds of Chinese character names to write Japanese phonetically. For example, the Japanese word *yama* 'mountain' was written phonetically by using the character sounding like *ya* with another character sounding like *ma*. Although there are earlier examples of the phonetic use of Chinese characters (such as in the songs of the *Kojiki* itself), it is known among Japanese grammarians as *man'yō-gana,* because its expression is most diversified in the *Man'yōshū.*

Two kinds of kana, or syllabic writing, developed from *man'yō-gana*. Katakana, which is angular in appearance, developed from the abbreviation of Chinese characters, and hiragana, rounded in appearance, by simplifying the grass (cursive) style of writing. Originally used as mnemonic symbols for reading Chinese characters, kana were eagerly adopted by women with literary aspirations; these women had been discouraged from learning Chinese characters, which belonged to the male domain of learning and writing. Murasaki Shikibu's 11th-century *Genji monogatari,* considered by many to be Japan's greatest literary achievement, was written almost entirely in hiragana. In contemporary Japanese writing, Chinese characters (kanji) and hiragana are used in combination, the former for content words and the latter for words such as particles and inflectional endings that indicate grammatical function. Katakana are used largely for foreign loanwords, telegrams, print advertising, and certain onomatopoeic expressions.

The use of kana made it possible to write a word in two ways. The Japanese word for 'mountain' could be written in kana (phonetically) by using two characters—that for *ya* and that for *ma*—or in kanji (by using the Chinese character meaning 'mountain'). This possibility helped to establish a relation between the Chinese character and its Japanese semantic equivalent and led to the practice of assigning a dual reading to Chinese characters: the Sino-Japanese reading (called *on-yomi*), based on the original Chinese pronunciation, and the Japanese reading (*kun-yomi*). Thus, the Chinese character originally meaning 'mountain' could be read as both *san* in *on-yomi* and *yama* in *kun-yomi*. Because Chinese words and their pronunciations were borrowed from different parts of China as well as during different historical periods, Modern Japanese includes many characters having more than one *on-yomi* reading.

The complexity of reproducing the strokes for each character and the multiple readings associated with it have stimulated movements to abolish Chinese characters in favour of kana writing or even more radical movements for completely romanizing the Japanese language. All these, however, have failed. Despite their complexity, Chinese characters retain a number of advantages over phonetic writing systems. For one thing, many homophonous words are visually distinguishable. For another, the meanings of unknown words written in Chinese characters can be surmised through the ideographic nature of these characters. This semantic transparency and the characteristic configurations of characters enable easy recognition and understanding of a passage. These strengths and Japan's high literacy rate make the abolishment of Chinese characters very unlikely.

Nevertheless, the shapes of Chinese characters have been simplified, and the number of commonly used characters has been limited. In 1946 the Japanese government issued a list of 1,850 characters for this purpose. Revised in 1981, the new list (called *Jōyō kanji hyō* "List of characters for daily use") contains 1,945 characters recommended for daily use. This basic list of Chinese characters is to be learned during primary and secondary education. When newspapers use characters not on the list, they also supply the reading in hiragana. (M.Si.)

AUSTRONESIAN LANGUAGES

General considerations

SIZE AND GEOGRAPHIC SCOPE

With approximately 1,200 members, the Austronesian language family includes about one-fifth of the world's languages. Only the Niger-Kordofanian family of Africa approaches it in number of languages, although both the Indo-European and Sino-Tibetan language families have considerably more speakers.

Before the European colonial expansions of the past five centuries, Austronesian languages were more widely distributed than any others, extending from Madagascar just off the southeast coast of Africa to Easter Island (Rapa Nui) some 2,200 miles west of Chile in South America—

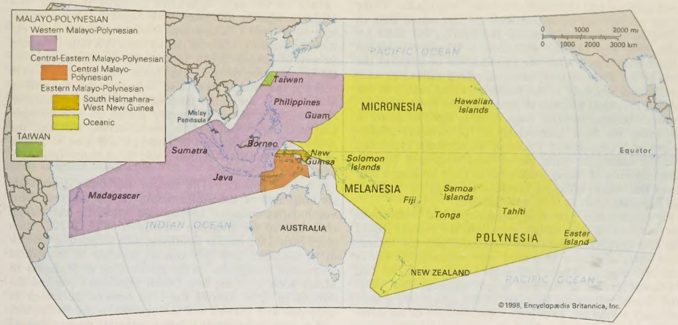


Figure 28: Major divisions of the Austronesian languages.

across an astonishing 206 degrees of longitude. Most of the languages are spoken within 10 degrees of the Equator, although some extend well beyond this, reaching as far north as 25° N latitude in northern Taiwan and as far south as 47° S latitude on New Zealand's South Island.

Despite the enormous geographic extension of the Austronesian languages, the relationship of many (though not all) of the languages can easily be determined by an inspection of such basic subsystems as personal pronouns or the numerals. Table 53 presents names for the numbers 1 to 10 in the Paiwan language of southeastern Taiwan, Cebuano Bisayan (Visayan) of the central Philippines, Javanese of western Indonesia, Malagasy of Madagascar, Arosi of the southeastern Solomon Islands in Melanesia, and Hawaiian.

Table 53: Numerals in Some Representative Austronesian Languages*

	Paiwan	Cebuano	Javanese	Malagasy	Arosi	Hawaiian
'one'	ita	usá	siji	isa	e-ta'i	e-kahi
'two'	ḍusa	duhá	loro	roa	e-rua	e-lua
'three'	tjelu	tulú	telu	telo	e-oru	e-kolu
'four'	sepatj	upát	papat	efatra	e-hai	e-hā
'five'	lima	limá	lima	dimy	e-rima	e-lima
'six'	unem	unúm	nem	enina	e-ono	e-ono
'seven'	pitju	pitú	pitu	fito	e-biu	e-hiku
'eight'	alu	walú	wolu	valo	e-waru	e-walu
'nine'	siva	siyám	sanga	sivy	e-siwa	e-iwa
'ten'	ta-puluq	púlu'	se-puluh	folo	ta-nga-huru	'umi

*Hyphens separate numerals from prefixed particles that are not shared by all languages.

Fourteen of the 21 or 22 Austronesian languages spoken by the pre-Chinese aboriginal population of Taiwan (also called Formosa) survive. Siraya and Favorlang, which are now extinct, are attested from fairly extensive religious texts compiled by missionaries during the Dutch occupation of southwestern Taiwan (1624–62). All the roughly 160 native languages of the Philippines are Austronesian, although it is likely that the now highly marginalized hunter-gatherer populations of Negritos originally spoke languages of other affiliations. Approximately 110 Austronesian languages are spoken in Malaysia, mostly in the Bornean states of Sabah and Sarawak. In mainland Southeast Asia some 7 or 8 Austronesian languages belonging to the close-knit Chamic group are spoken in Vietnam, in Cambodia, in border regions of Laos, and on Hainan Island in southern China.

The language of Madagascar

Malagasy generally is regarded as a single language, although it may have as many as 20 dialects, some of which approach the dialect-language limit. The remaining 900 Austronesian languages are about equally divided among Indonesia (including Irian Jaya, the western half of the large island of New Guinea) and the Pacific islands of Melanesia, Micronesia, and Polynesia. The great majority of Austronesian languages in the Pacific are found in Melanesia, particularly in coastal areas of New Guinea and the islands of the Bismarck Archipelago (New Britain, New Ireland, the Admiralty Islands). The Austronesian languages of Melanesia are often found closely interspersed with an older population of non-Austronesian languages, collectively known as Papuan. With few exceptions the Austronesian languages of Melanesia tend to be spoken in coastal areas and on small offshore islands.

MAJOR LANGUAGES

Major Austronesian languages include Cebuano, Tagalog, Ilokano, Hiligaynon, Bikol, Waray, Kapampangan, and Pangasinan of the Philippines; Malay, Javanese, Sundanese, Madurese, Minangkabau, the Batak languages, Acehnese, Balinese, and Buginese of western Indonesia; and Malagasy of Madagascar. Each of these languages has more than one million speakers. Javanese alone accounts for about one-quarter of all speakers of Austronesian languages. In eastern Indonesia the average number of speakers per language drops to a few tens of thousands and in western Melanesia to fewer than a thousand. In the central Pacific, where the average number of speakers per language again increases to more than 100,000, the major languages include Fijian, Samoan, and Tongan.

Tagalog forms the basis of Pilipino, the national language of the Philippines, and the Merina dialect of Malagasy, which is spoken in the highlands around the capital of Antananarivo, forms the basis for standard Malagasy. Hindu-Buddhist polities, based on Indian concepts of the state, arose in parts of the Malay Peninsula and Sumatra during the first few centuries of the Christian era and somewhat later in Java. As a result of these contact influences, Sanskrit loanwords entered Malay and Javanese in large numbers. Many Philippine languages also contain substantial numbers of Sanskrit loans, even though no part of the Philippines was ever Indianized. It is generally agreed that these and the later Arabic and Persian loanwords that are found in Philippine languages were transmitted through the medium of Malay.

It is now widely agreed, following the pioneering thesis of the Norwegian linguist Otto Christian Dahl, that Madagascar was settled by immigrants from southeastern Borneo sometime between the 7th and 13th centuries AD. The presence of Sanskrit loans in Malagasy suggests that the movement to Madagascar took place after the beginnings of Indianization in western Indonesia, while the presence of some Arabic loans that show distinctive Malay adaptations suggests that contact between Madagascar and Malay-speaking portions of western Indonesia may have continued after the initial migration from Southeast Asia.

Of all Austronesian languages, Malay—which is native to the Malay Peninsula, adjacent portions of southern and central Sumatra, and some smaller neighbouring islands—

probably has had the greatest political importance. Three stone inscriptions associated with the Indianized state of Śrivijaya in southern Sumatra and bearing the dates AD 683, 684, and 686 are written in a language generally called Old Malay. After the introduction of Islām at the end of the 13th century, Malay-speaking sultanates were established not only in the Malay-speaking region of the Malay Peninsula but also in Brunei on the coast of northwestern Borneo. In other areas, such as Aceh of northern Sumatra, the Sulu Archipelago of the southern Philippines, and Ternate and Tidore of the northern Moluccas, Islāmic sultanates made use of local languages, but the large number of Malay loanwords in these languages suggests that Malay-speaking missionaries must have played an important part in their establishment.

Malay and the introduction of Islām

Fairly abundant palm-leaf manuscripts and inscriptions on stone or various metals constitute the textual record for Old Javanese, a language associated with the Indianized states of eastern Java from approximately the 9th to the 15th century. About half of the vocabulary of the Old Javanese texts is of Sanskrit origin, although this material clearly reflects the language of the courts and almost certainly would not have been representative of the common people.

The historical importance of both Tagalog and Malay probably was favoured by geographic considerations. Tagalog is the language native to the region of Manila Bay. When the Spanish initiated the 350-year-long Manila galleon trade in 1565 they found a preexisting trade network linking Fukienese traders from southern China with the local native population and probably with some Malay traders from western Indonesia. Malay was spoken on both sides of the strategic Strait of Malacca between Sumatra and the Malay Peninsula. When the India-China trade commenced at approximately the start of the 1st century AD, the favoured sea route passed through the Strait of Malacca, drawing the Malay-speaking populations of this region into a much wider network of international commerce. When representatives of the Dutch East India Company arrived in Indonesia at the beginning of the 17th century, they discovered that Malay served as a lingua franca in major ports throughout the archipelago; the language has retained that role to the present day. It was thus natural that Malay would be selected as the basis for the national language of Malaysia (Bahasa Malaysia), Brunei (Bahasa Kebangsaan 'national language'), and Indonesia (Bahasa Indonesia). In Indonesia speakers of Malay were far outnumbered by speakers of Javanese, but there Malay offered a neutral alternative to the widely perceived threat of ethnic domination by the overwhelming Javanese-speaking majority.

A similar geographic determinism favouring the rise of local languages to the status of lingua francas can be seen on a smaller scale in Melanesia. Motu, centred in the important harbour of Port Moresby in Papua New Guinea, was the medium through which the seasonal *hiri* (trading voyages) took place across the 225-mile-wide Gulf of Papua before the arrival of Europeans. Under British colonial rule a simplified form of Motu known as Hiri, or Police, Motu served as the language of the territorial constabulary. Tolai, spoken natively around the important harbour town of Rabaul on the island of New Britain, came under heavy contact influence from English in a 19th-century plantation setting. The result was a creolized form of the language known as Melanesian Pidgin, or Tok Pisin, today one of the national languages of Papua New Guinea.

Hiri, or Police, Motu

WRITTEN DOCUMENTS

Pre-19th century. *Pre-16th century.* The earliest written documents in an Austronesian language are three Old Malay inscriptions from southern Sumatra dating to the late 7th century. The earliest dated inscription in Cham, the language of the Indianized kingdom of Champa in central Vietnam, bears a date of AD 829, although some undated inscriptions may be older. An Old Malay stone inscription from central Java is dated to AD 832 and attests to the high prestige of Malay in areas where it was not a native language.

Much of the early epigraphic material in Cham and Malay is heavily interlaced with Sanskrit, and some inscriptions from Champa and southern Sumatra are entirely in Sanskrit. Material dating from this time is written in any of several South Indian scripts. Sometime after the introduction of Islām and before the end of the 13th century, the Arabic script also came into use for writing Malay and a few other languages of western Indonesia. At the end of the 20th century almost all Austronesian languages were written in a roman script, although the Arabic script (called Jawi in Malay) is still used in certain contexts in Malay, Acehnese, and some other languages of western Indonesia.

16th–18th century. The earliest European documents on languages of the Austronesian family are two short vocabularies collected by Antonio Pigafetta, the Italian chronicler of the Magellan expedition of 1519–22. Dutch ships bound for insular Southeast Asia stopped to restock in Madagascar, and this contact resulted in an almost immediate recognition of the relationship of Malagasy to Malay soon after the first Dutch expedition reached Indonesia in 1596. During the 17th century the Dutch in Indonesia and Taiwan and the Spanish in the Philippines and Guam compiled the first substantial descriptions of Austronesian languages.

By the beginning of the 18th century the Dutch scholar Hadrian Reland was able to suggest an eastward extension of Malay-like languages into the western Pacific. Following the three Pacific voyages of James Cook from 1768 to 1780, the close similarity of the Polynesian languages to one another—and their more general similarity to Malay—became widely known, although it was mistakenly believed, largely on racial grounds, that the languages of Melanesia were not related to those of Polynesia or to one another.

19th–20th century. *Early classification work.* By 1834 the British historian and linguist William Marsden was able to speak of languages such as Malagasy and Malay as Hither Polynesian and of the languages of the central and eastern Pacific as Further Polynesian, although he offered no name for the language family as a whole. The German scholar Wilhelm von Humboldt is generally credited with coining the name Malayo-Polynesian, although the word first appeared in print in an 1841 publication of his contemporary, the German linguist Franz Bopp. Several decades later Robert Codrington, a leading English scholar of the languages of Melanesia, objected to the designation Malayo-Polynesian on the grounds that it excludes the darker-skinned peoples of Melanesia. He referred instead to the "Ocean" family of languages. In 1906 the Austrian anthropologist and linguist Wilhelm Schmidt proposed that the Munda languages of eastern India and the Mon-Khmer languages of mainland Southeast Asia form a language family, which he christened Austroasiatic (meaning "southern Asian"). Primarily on the basis of similarities in verbal affixes, Schmidt further suggested that the Malayo-Polynesian languages and the Austroasiatic languages form a superfamily that he designated Austric. In accordance with his newly coined terminology he substituted Austronesian ("southern islands") for the older family name. Both names were used extensively in the 20th century, although since the mid-1960s the name Malayo-Polynesian has been restricted to various large subgroups of Austronesian rather than applied to the language family as a whole.

Wilhelm von Humboldt

The first analysis of Austronesian languages to make use of the comparative method of linguistics is attributed to the Dutch-Indonesian scholar H.N. van der Tuuk, whose comparisons during the 1860s and '70s showed that various languages in the Philippines and Indonesia could be related to a common ancestor through recurrent similarities in the forms of words. Van der Tuuk's central achievement in comparative linguistics was the establishment of what later came to be known as the RGH law, or van der Tuuk's first law; it describes the recurrent sound correspondence of Malay /r/ to Tagalog /g/ and Ngaju Dayak /h/, as in Malay *urat,* which corresponds to Tagalog *ugat* and Ngaju Dayak *uhat* 'vein.' In addition, van der Tuuk's grammar of the Toba Batak language of northern Sumatra, published in two volumes between 1864 and 1867, stands as one of the earliest attempts to represent a non-Western language in terms of inductively derived categories rather than in terms of traditional Latin grammar. Despite his many achievements, however, van der Tuuk's work included only languages in Indonesia and the Philippines. In the 1880s the Dutch Sanskrit scholar Hendrik Kern began a series of studies that in principle encompassed the entire Austronesian family, drawing on data from both island Southeast Asia and the Pacific. The first true systematizer in the Austronesian field was the Swiss scholar Renward Brandstetter, whose work in the period 1906–15 led to the reconstruction of a complete sound system for what he called Original Indonesian and the compilation of a very preliminary comparative dictionary. Like van der Tuuk, however, Brandstetter worked only on the Austronesian languages of island Southeast Asia.

The work of Otto Dempwolff. The modern study of the Austronesian languages is generally traced to the German medical doctor and linguist Otto Dempwolff, whose three-volume *Comparative Phonology of Austronesian Word Lists,* published between 1934 and 1938, established a more complete sound system than that of Brandstetter and further took account of languages in all the major geographic regions rather than just insular Southeast Asia.

Dempwolff's comparative dictionary

Dempwolff also published the first comprehensive comparative dictionary of Austronesian languages, with some 2,200 reconstructed words based on evidence from 11 modern languages: Tagalog, Toba Batak, Javanese, Ngaju Dayak, Malay, and Malagasy (which he called Indonesian languages); Sa'a and Fijian (called Melanesian languages); and Tongan, Futunan, and Samoan (called Polynesian languages). Although Dempwolff's phonological reconstruction has undergone considerable revision, especially in light of evidence from the aboriginal languages of Taiwan, and although his comparative dictionary is now very much out of date, his work remains the foundation for much of what has followed.

Classification and prehistory

MAJOR SUBGROUPS

Given the size of the Austronesian family, the subgrouping of the languages is a matter of some importance, bearing on, among other things, the determination of the Austronesian homeland. Until the 1930s the branches of Austronesian were customarily identified with purely geographic labels: Indonesian, Melanesian, Micronesian, and Polynesian. The inadequacy of this subdivision is apparent; Polynesian, for example, is known to encompass not only the languages of Polynesia but also Polynesian Outlier languages of both Melanesia and Micronesia. Moreover, each of the other geographically defined groups turns out to be a heterogeneous collection of languages that belong to more than one linguistically defined group.

The first breakthrough in the subgrouping of the Austronesian languages was made by Dempwolff in the second volume of his distinguished trilogy, where he concluded that the languages of Polynesia and most of those of Melanesia and Micronesia share a number of innovative features that are most plausibly attributed to changes in a single protolanguage, which he named Urmelanesisch (Proto-Melanesian) and which is known today as Proto-Oceanic. The Oceanic hypothesis maintains that all Austronesian languages east of a line that runs through Indonesian New Guinea (Irian Jaya) at approximately 138° E longitude—except for Palauan and Chamorro of western Micronesia—are descended from a single protolanguage spoken many generations after the initial breakup of Proto-Austronesian itself.

The major subgroups of Austronesian as generally recognized today are described below.

The highly diverse Formosan group

Formosan. The term Formosan language is not to be understood as representing a subgroup defined by exclusively shared innovations. Rather, it is a collective term for a highly diverse collection of languages, most of which share broad typological similarities with languages in the Philippines and some other areas (such as Madagascar). The Yami language, which is spoken on Lan-yü (Botel Tobago) island off the southeastern coast of Taiwan, forms a subgroup with Ivatan and Itbayaten in the northern Philippines. The other 14 surviving aboriginal languages of Taiwan may fall into as many as six primary branches of the language family, each one coordinate with the entire Malayo-Polynesian branch. Under such circumstances very small subgroups or even single languages provide an independent line of evidence for the nature of Proto-Austronesian that is theoretically equivalent to the entire Malayo-Polynesian branch of some 1,180 member languages. Among the best-described Formosan languages are Atayal (spoken in the northern mountains), Amis (spoken along the narrow east coast), and Paiwan (spoken near the southern tip of the island); only superficial descriptions are available for most of the other Formosan languages.

Western Malayo-Polynesian (WMP). Although Western Malayo-Polynesian is a convenient cover term for the Austronesian languages of the Philippines, western Indonesia (Borneo, Sumatra, Java-Bali-Lombok, Sulawesi), mainland Southeast Asia, Madagascar, and at least Chamorro and Palauan in western Micronesia, it is in effect a catchall category for the Malayo-Polynesian languages that do not exhibit any of the innovations characteristic of Central-Eastern Malayo-Polynesian and may very well contain several primary branches of Malayo-Polynesian. As mentioned previously, some of the largest and best-known Austronesian languages—including Ilokano, Tagalog, Cebuano, Malay, Acehnese, Toba Batak, Minangkabau, Sundanese, Javanese, Balinese, Buginese, Makasarese, and Malagasy—are Western Malayo-Polynesian.

Central Malayo-Polynesian (CMP). The Central Malayo-Polynesian languages are found throughout much of eastern Indonesia, including the Lesser Sunda Islands from Sumbawa through Timor, and most of the Moluccas. Many of the changes that define this linguistic group cover most of the languages but do not reach the geographic extremes, and the group has therefore been questioned by some scholars. Few of the languages are large or well-known, but those for which fuller descriptions are available include Manggarai and Ngadha, spoken on the island of Flores; Roti, spoken on the island of the same name; Tetum, spoken on the island of Timor; and Buruese, spoken on the island of Buru in the central Moluccas.

South Halmahera–West New Guinea (SHWNG). This small group of Austronesian languages is found in the northern Moluccan island of Halmahera and in the Doberai Peninsula (also called Vogelkop or Bird's Head) of western New Guinea. Preliminary descriptions exist only for Buli of Halmahera and Numfor-Biak and Waropen of western New Guinea; most of the languages are known only from short word lists.

The largest major subgroup

Oceanic (OC). The Oceanic subgroup is the largest and best defined of all major subgroups in Austronesian. It includes all the languages of Polynesia, all the languages of Micronesia (except Palauan and Chamorro), and all the Austronesian languages of Melanesia east of the Mamberamo River in Irian Jaya. Some of the better-known Oceanic languages are Motu of southeastern New Guinea, Tolai of New Britain, Sa'a of the southeastern Solomons, Mota of the Banks Islands in northern Vanuatu, Chuukese (Trukese) of Micronesia, Fijian, and many Polynesian languages, including Tongan, Samoan, Tahitian, Maori, and Hawaiian. Yapese, long considered unplaceable, now appears to be Oceanic, although its place within Oceanic remains obscure.

LOWER-LEVEL SUBGROUPS

Philippine languages. One of several identifiable lower-level units within these major subgroups is the Philippine group within Western Malayo-Polynesian. It consists of Yami, spoken on Lan-yü (Botel Tobago) island off the southeastern coast of Taiwan; almost all the languages of the Philippine Islands; and the Sangiric, Minahasan, and Gorontalic languages of northern Sulawesi in central Indonesia. The Samalan dialects—spoken by the Sama-Bajau, the so-called sea gypsies in the Sulu Archipelago, and elsewhere in the Philippines—do not appear to belong to the Philippine group, and their exact linguistic position within the Austronesian family remains to be determined. Although the term Philippine language or Philippine-type language has been applied to such languages as Chamorro of the Mariana Islands or the languages of Sabah in northern Borneo, this label is typological rather than genetic.

Polynesian languages. Perhaps the best-known lower-level subgroup of Austronesian languages is Polynesian, which is remarkable for its wide geographic spread yet close relationship. The "Polynesian triangle," defined by Hawaii, Easter Island, and New Zealand, encloses Polynesia proper, an area about twice the size of the continental United States. In addition, some 18 Polynesian-speaking societies, the above-mentioned Polynesian Outliers, are found in Micronesia and Melanesia.

The Polynesian languages generally are divided into two branches, Tongic (Tongan and Niue) and Nuclear Polynesian (the rest). Nuclear Polynesian in turn contains Samoic-Outlier and Eastern Polynesian. Maori and Hawaiian, two Eastern Polynesian languages that are separated by some 5,000 miles of sea, appear to be about as closely related as Dutch and German. The closest external relatives of the Polynesian languages are Fijian and Rotuman, a non-Polynesian language spoken by a physically Polynesian population on the small volcanic island of Rotuma northwest of the main Fijian island of Viti Levu; together with Polynesian, Fijian and Rotuman form a Central Pacific group. A number of proposals have been made regarding the immediate relationships of the Central Pacific languages; the majority of these suggest a grouping of Central Pacific with certain languages in central and northern Vanuatu, but these proposals remain controversial.

Nuclear Micronesian. Most of the languages of Micronesia are Oceanic, and, with the possible exception of Nauruan, which is still poorly described, they form a fairly close-knit subgroup that is often called Nuclear Micronesian. Palauan, Chamorro (Mariana Islands), and Yapese (western Micronesia) are not Nuclear Micronesian languages; the former two appear to be products of quite distinct migrations out of Indonesia or the Philippines, and, while Yapese probably is Oceanic, it has a complex history of borrowing and does not readily seem to form a subgroup with any other language.

Aberrant languages. Yapese is one of several problematic languages that can be shown to be Austronesian but that share little vocabulary with more typical languages. Other languages of this category are Enggano, spoken on a small island of the same name situated off the southwest coast of Sumatra, and a number of Melanesian languages.

Shifting classification

In the most extreme cases the classification of a language as Austronesian or non-Austronesian has shifted back and forth repeatedly, as with the Maisin language of southeastern Papua New Guinea (now generally regarded as an Austronesian language with heavy contact influence from Papuan languages). Other controversial or aberrant languages are Arove, Lamogai, and Kaulong of New Britain, Ririo and some other languages of the western Solomons, Asumboa of the Santa Cruz archipelago, Aneityum and some other languages of southern Vanuatu, several languages of New Caledonia, and Nengone and Dehu of the Loyalty Islands in southern Melanesia. Atayal of northern Taiwan is an example of a language once considered to be highly aberrant in vocabulary, but it is much less distinctive now that researchers have found that the Squliq dialect (which was chosen as representative of Atayal) exhibits idiosyncratic changes owing to a historical form of "speech disguise" characteristic of men's speech. This feature is still preserved in the Mayrinax dialect of the Cʔuliʔ dialect cluster.

PREHISTORIC INFERENCES FROM SUBGROUPING

The view, current from roughly 1965 to 1975, that Melanesia is the area of greatest linguistic diversity in Austronesian and that the Austronesian homeland therefore must have been in Melanesia has been shown to be inconsistent both with the comparative method of linguistics and with archaeological indications that Austronesian speakers entered the western Pacific from island Southeast Asia about 2000 BC. It has accordingly been abandoned by virtually all scholars.

Both linguistic and archaeological evidence point to an initial dispersal of Austronesian languages from Taiwan several centuries after Neolithic settlers introduced grain agriculture, pottery making, and domesticated animals to the island from the adjacent mainland of China about 4000 BC. By perhaps 3500 BC, populations bearing a clear cultural resemblance to those in Taiwan had begun to appear in the northern Philippines, and within a millennium similar material traces appear throughout Indonesia. The linguistic evidence suggests a steady southward and eastward movement, with Austronesian speakers moving around the northern coast of New Guinea into the western Pacific about 2000 BC. From the region of New Guinea and the Bismarck Archipelago settlers fanned out very rapidly, crossing the sea with highly seaworthy outrigger canoes. In Oceania the dispersal of Austronesian-speaking peoples is most closely associated archaeologically with the distribution of Lapita pottery. Because the earliest Lapita sites in Fiji and western Polynesia are only three or four centuries younger than the earliest dated Lapita site in western Melanesia, the colonization of Melanesia as far east as Fiji appears to have been accomplished within 15 or 20 generations. There is a puzzling thousand-year gap before the settlement of central and eastern Polynesia, with Hawaii being settled only within the past 1,500–1,700 years and New Zealand within roughly the past millennium.

The settlement of Micronesia

The settlement history of Micronesia is more complex: Palau and the Mariana Islands were settled by two migrations which were distinct from that associated with Lapita pottery. Most of the low coral atolls of the Caroline Islands were settled by 2000 BP, but some radiocarbon dates from the Marshall Islands suggest that Austronesian speakers may have reached the atolls of Micronesia not long after the settlement of Fiji and western Polynesia.

EXTERNAL RELATIONSHIPS

Speculation concerning the external relationships of Austronesian languages has ranged far and wide. In the first half of the 19th century Bopp, who was a distinguished Indo-Europeanist, became convinced of the relationship of Indo-European to Austronesian. This theme was taken up again in the 1930s by Brandstetter. In 1942 the American linguist Paul K. Benedict initiated the Austro-Tai hypothesis, a proposed connection between the Tai languages and various minority (Kadai) languages on the mainland of Southeast Asia. Other researchers have proposed connections with Japanese (as has Benedict himself), the Papuan languages of New Guinea, various American Indian languages, Chinese, and Ainu. In short, almost every language family that might conceivably be related to Austronesian simply on grounds of a priori geographic proximity has been proposed as a relative, the one notable exception to date being Australian Aboriginal languages. Most of these proposals are speculative and have not achieved a general following.

Benedict's Austro-Tai hypothesis has perhaps received the widest attention in recent years, as it has been advocated in a large number of publications. However, in some ways the most compelling hypothesis for a wider language grouping that includes Austronesian is the Austric hypothesis, linking the Austroasiatic languages (the Munda languages of eastern India and the Mon-Khmer languages of mainland Southeast Asia) with Austronesian. The original hypothesis, first proposed in 1906 by Wilhelm Schmidt and long neglected by most linguists, has been greatly strengthened by more recent research.

Structural characteristics of Austronesian languages

SYNTAX

Word order. Although some linguists have questioned the usefulness of the notion of subject in Philippine languages, it remains a pivotal concept in typological studies of word order. The great majority of Formosan and Philippine languages are verb–subject–object (VSO) or VOS. This statement is true of virtually all the Formosan languages, with the minor qualification that auxiliaries and markers of negation may precede the main verb. Some contemporary languages, such as Thao and Saisiyat, have

SVO word order, but there are indications that this is a relatively recent adaptation to the similar word order of Taiwanese, the Chinese language with which the Formosan languages have been in longest contact.

The presence of VSO languages

Most languages of western Indonesia—such as Malay, Javanese, or Balinese—are SVO. However, a smaller number of languages, including Malagasy, the Batak languages of northern Sumatra, and Old Javanese (as opposed to modern Javanese), begin sentences with a verb. The majority of Austronesian languages in both eastern Indonesia and the Pacific are also SVO. The major exceptions to this pattern are in coastal areas of New Guinea, where a number of Austronesian languages are SOV, and the Polynesian languages and Fijian, which are VSO. The SOV languages of New Guinea also exhibit other features universally characteristic of verb-final languages, such as the use of postpositions (*e.g.,* "the house in") rather than prepositions ("in the house"). It is generally agreed that these Austronesian languages evolved to their present state as a result of generations of contact with Papuan languages, which typically are SOV.

Verb systems. Perhaps the most fundamental distinction in the verb systems of Austronesian languages is the division into stative and dynamic verbs. Stative verbs often translate as adjectives in English, and in many Austronesian languages it is doubtful whether a category of true adjectives exists. Examples of stative verbs are 'to be afraid,' 'to be sick/painful,' 'to be new,' 'to sleep/to be asleep,' and colour words. In some languages the stative prefix *ma-* can be added to higher numerals, as in Maranao *ma-gatos* 'one hundred.'

Dynamic verbs generally are more complex than stative verbs. Most Formosan and Philippine languages and many of the languages of Sulawesi have a large inventory of affixes used to create different nuances of meaning in verbal or nominal stems. Most noteworthy is the system of verbal focus, which has been the centre of controversy and the subject of many conflicting interpretations since 1917, when Leonard Bloomfield provided the first detailed description of Tagalog syntax. The major verbal focuses of Tagalog can be illustrated as follows:

A sentence that focuses on the actor (subject) is marked by *-um-*; for example, *b-um-ilí ang lalake ng tinapay sa tindahan* 'the man bought some bread at the store' (literally, 'buy *ang* man *ng* bread *sa* store') or *b-um-ilí si Maria ng tinapay sa tindahan* 'Maria is buying/bought some bread at the store' (literally, 'buy *si* Maria *ng* bread *sa* store'). A sentence that focuses on the patient (object) is marked by *-in-* in the past, and by *-in* in the nonpast); for example, *b-in-ilí ni Maria ang tinapay sa tindahan* 'Maria bought the bread at a/the store' (literally, 'bought *ni* Maria *ang* bread *sa* store') or *bilh-ín ni Maria ang tinapay sa tindahan* 'Maria is buying the bread at a/the store.' A sentence that has a locative focus is marked by *-an*; for example, *b-in-ilh-ân ng babae ng tinapay ang tindahan ni Aling Maria* 'the woman bought some bread at Maria's store' (literally, 'bought *ng* woman *ng* bread *ang* store'). A sentence with an instrumental or benefactive focus is marked by *i-*; for example, *i-b-in-ilí ni Maria ng tinapay ang pera nang tatay-niyá* 'Maria bought some bread with her father's money' or *i-b-in-ilí ni Maria ng tinapay si Juan* 'Maria bought (some) bread for Juan.'

The focus affix

In each of the above sentences one noun is marked as being in focus. Focused personal nouns (proper names or common nouns that can be used as proper names, such as 'Mother' or 'Father') are preceded by *si.* Focused common nouns are preceded by *ang,* and the combination is commonly called the "*ang*-phrase." The syntactic relationship that the focused noun bears to the verb is signaled by the focus affix (*e.g.,* actor, patient). Moreover, focused noun phrases are definite, or old information, while nonfocused noun phrases may be either definite or indefinite. The speaker's choice of focus thus depends to a large extent on discourse factors. Similar systems of encoding syntactic relationships are widespread in Formosan and Philippine languages, in the languages of Sabah (formerly North Borneo), in those of northern Sulawesi (northern Celebes), in the Chamorro language of western Micronesia, and in Malagasy. Somewhat less similar systems with some of the same features are found in the Batak languages of northern Sumatera (northern Sumatra) and in Old Javanese.

One school holds that focus is voice. Under this interpretation such languages as Tagalog have only one active voice but three types of passives: a direct passive, a local passive, and an instrumental or benefactive passive. A second school holds that focus is case-marking: the case roles of subjects are marked by the focus affix on the verb. What distinguishes focus systems from the simple active-passive voice systems of such languages as Malay or modern Javanese is their ability by means of verbal affixation to express prepositional phrases as subjects. When the prepositional phrase is not in focus it is expressed as a preposition followed by a noun rather than as an *ang*-phrase: compare the third example above, *b-in-ilh-ân ng babae ng tinapay ang tindahan* 'the woman bought the bread at the store,' where *ang tindahan* 'the store' is in focus and the locative relationship is expressed by the verb suffix *-an,* with any of the other sentences that contain *tindahan* 'store,' where the locative relationship is expressed by the preposition *sa.*

One feature of the verb systems of many Austronesian languages is particularly noteworthy: nonsubject actors and possessors are marked in the same way (in Tagalog these are marked with the particle *ni*). As a result 'was bitten by the dog' and 'the dog's biting (of something)' have identical structures. Because of this ambiguity the focus affixes in most focus languages create both verbs and nouns. Where focus has been lost, as in much of Indonesia and the Pacific, the remnant affixes may be used only to create nouns.

Pronouns. Almost all Austronesian languages distinguish two forms of 'we': an inclusive form (listener included) and an exclusive form (listener excluded). Many languages in the Philippines have a special dual inclusive ('you and me'). In addition to singular and plural numbers, some Oceanic languages distinguish a dual number ('we two,' 'you two,' 'the two of them'). A few Oceanic languages distinguish a fourth number that is greater than two but smaller than a typical plural. Historically this number derives from the Proto-Austronesian word for 'three,' but it may in fact apply to numbers up to five and so is sometimes called "paucal" ('a few'). Gender is rarely if ever distinguished. Tables 54 and 55 illustrate two fairly typical pronoun systems, one from Tagalog, a Western Malayo-Polynesian language, and the other from Fijian, an Oceanic language.

Absence of gender indicator

Table 54: Pronouns in Tagalog

	first person (exclusive)	first person (inclusive)	second person	third person
Singular	akó	—	ikáw/ka	siyá
Dual	—	katá	—	—
Plural	kamí	tayó	kayó	silá

Table 55: Pronouns in Fijian

	first person (exclusive)	first person (inclusive)	second person	third person
Singular	au	—	o	e
Dual	keirau	(e)daru	(o)drau	(e)rau
Paucal	keitou	([e]da)tou	(o)dou	(e)ratou
Plural	keimami	(e)da	(o)nī	(e)ra

Neither of these sets exhausts the pronouns available in the languages shown. Tagalog has distinct sets of pronouns for *ang, ng,* and *sa* grammatical slots, and only the *ang* set is shown here; Fijian has distinct subject and object sets, and only the subject set is shown here.

Probably the most spectacular pronominal feature in Austronesian languages is the expression of possessive-marking in Oceanic languages. In many of the languages of Melanesia, nouns are marked for one of two types of possessive relationship, generally called "inalienable" and "alienable." Inalienable categories include body parts, certain kinship relationships, and such "spiritual" aspects of an individual as his shadow (often associated with the soul) and his name. Inalienable possession is marked by suffixing a possessive pronoun to the possessed noun, as in Fijian *na mata-na* 'his eye' (literally, '[article] eye-his') or *na tama-qu* 'my father.' Alienable possession is expressed by suffixing the possessive pronoun to a generally preposed classifying particle that specifies any of several possible relationships between the possessed noun and the possessor, as in Fijian *na no-na vale* 'his house' (literally, '[article] neutral-his house'), *na ke-na ika* 'his fish (to eat)' ('[article] edible-his fish'), and *na me-na dovu* 'his sugarcane (to suck the juice from)' ('[article] drinkable-his sugarcane'). The distinction between neutral and edible possession is widespread in Oceanic languages, and it appears in a few languages of eastern Indonesia. The further distinction of drinkable possession has a more limited distribution.

The Polynesian languages have a somewhat different system of possessive marking. The most prominent feature of this system is the contrast between what are sometimes called "dominant" and "subordinate" possession. In dominant possession the possessor generally has a relationship of control, as with Hawaiian *ka ki'i a Lani* 'the picture taken or painted by Lani,' while in subordinate possession this sense of control does not exist, as in *ka ki'i o Lani* 'the picture taken or painted of Lani.'

Numbers and number classifiers. As illustrated in Table 53, most Austronesian languages have a decimal system of counting. Others, such as Ilongot of the northern Philippines and some of the languages of the Lesser Sunda Islands in eastern Indonesia, have quinary systems (*i.e.*, systems based on five). In the New Guinea area several Austronesian languages have radically restructured number systems that probably result from intensive contact with neighbouring Papuan languages. An example is Gapapaiwa of Milne Bay, with *sago* 'one,' *rua* 'two,' *rua ma sago* 'three' (literally, 'two and one'), *rua ma rua* 'four,' *rua ma rua ma sago* 'five.' In such systems counting is often limited to relatively small quantities.

The counting of objects

A number of the languages of Indonesia and the Pacific use number classifiers in counting objects, as with Bahasa Indonesia *se-buah rumah* 'a house' (literally, 'one-fruit house'), *se-orang guru* 'a teacher' (literally, 'one-person teacher'), or *se-batang rokok* 'a cigarette' (literally, 'one-trunk cigarette'). In some languages of Micronesia the traditional counting systems were highly complex, with upwards of 30 number classifiers that distinguished counted objects by their shape, animateness, and other features.

Spacial orientation. Some Austronesian languages have terms for the cardinal directions east, west, north, and south, but in most cases these appear to have developed after European contact and may sometimes be due to inaccurate reporting by Europeans.

The system of directional orientation found in many Austronesian languages is constructed on two axes, a land-sea axis and a monsoon axis. The land-sea axis is very widespread among Austronesian-speaking peoples. Two widely separated examples are Thao (central Taiwan) *tana-saya* 'uphill, toward the mountains,' *tana-raus* 'downhill, toward the sea' and Hawaiian *mauka* 'toward the mountains,' *makai* 'toward the sea.' The monsoon axis is geographically more restricted, but the earlier reconstructed terms **habaRat* 'west monsoon' and **timuR* 'southeast monsoon' have been preserved in languages outside the monsoon region, though with change of meaning (*e.g.*, Samoan *afā* 'storm, gale, hurricane,' *timu* 'be rainy').

Demonstrative pronouns often distinguish two forms of 'there.' In some languages these correspond to second-person and third-person pronominal reference: 'there (near the listener)' versus 'there (near a third person).' In other languages a distinction is made between a referent that is visible versus a referent that is not visible.

MORPHOLOGY AND CANONICAL SHAPE

Verb morphology. The Austronesian languages of Taiwan, the Philippines, northern Borneo, and Sulawesi and some other languages (such as Malagasy, Palauan, and Chamorro) are characterized by a very rich morphology. Some languages use affixation to encode many types of syntactic relationships that are expressed in most other languages through the use of free words. Thao of central Taiwan, for example, allows aspect markers to be attached to prepositional phrases, as in *in-i-nay yaku* 'I was here' (literally, '[past]-location-this I'). In Thao, relative clauses are expressed through attributive constructions that may use complex nouns derived by affixation, as in *m-ihu a s-in-aran-an yanan sapaz* 'the place where you walked has footprints' ('your [ligature-past]-walking-place has footprints'). Most of the so-called focus affixes in such languages have both verbalizing and nominalizing functions.

Prefixed subject markers on the verb

Many of the languages of Sulawesi and eastern Indonesia have prefixed subject markers on the verb. In some languages these co-occur with full free pronouns marking the subject and so function like a system of agreement. In some of the languages of western Melanesia, such as Motu, the verb complex consists of a prefixed subject marker, the verb stem, and a suffixed object marker, together with free nouns or pronouns marking subject and object, producing structures such as 'the man the dog he-kicked-it' for 'the man kicked the dog.' In a case such as this, the structure of the verb complex provides a clue that the current SOV order of sentence constituents has developed from an earlier SVO order.

Reduplication. Reduplication takes numerous forms and has a great variety of functions in Austronesian languages. Partial reduplication of a verb stem is used to mark the future tense in both Rukai of Taiwan and Tagalog of the Philippines, as in Tagalog *l-um-akad* 'walk' but *la-lakad* 'will walk' or *s-um-ulat* 'write,' *su-sulat* 'will write.' Full reduplication is used to mark plurality of nouns in Bahasa Indonesia, as with *anak* 'child' but *anak anak* 'children.' In many languages reduplication is used together with affixation to express a variety of semantic nuances. The pattern seen in Indonesian *anak anak-an* 'doll' or *orang orang-an* 'scarecrow' (*orang* 'person') is only one of many that occur in various languages.

Submorphemes. Linguists have generally maintained that the smallest meaning-bearing units of language structure are morphemes, elements that are isolated by the contrast of partially similar words, as in *berry* : *cranberry* (hence both *cran* and *berry* are morphemes of English). However, English words such as *glow, glimmer, glisten, glitter, glare, glint, gloss,* and the like exhibit a recurrent association of sound and meaning without contrast. Many Austronesian languages, particularly in insular Southeast Asia, show similar types of recurrent sound-meaning associations that are not defined by contrast. In the great majority of cases, these consist of the last syllable of a morpheme. A clear illustration is seen in Malay, where about 40 two-syllable words end in *-pit* and roughly half of these have meanings that can be characterized as referring to the approximation of two surfaces, as in *(h)apit* 'pressure between two disconnected surfaces,' *capit* 'pincers,' *men-cepit* 'to nip,' *dempit* 'pressed together, in contact,' *gapit* 'nipper, clamp,' and *kempit* 'carry under the arm.'

Canonical shape. The term canonical shape refers to the clearly marked preferences that some languages show for number of syllables, sequencing of consonants and vowels, and so on in the construction of words. Many Austronesian languages show a clear preference for a disyllabic (two-syllable) canonical shape in content words (words that have a reference rather than a purely grammatical function). Where this preference is violated by the operation of other forces, it often reasserts itself through special mechanisms. Javanese *əri* 'thorn' passed through a stage in which it was *ri* but gained a schwa to meet the preferred two-syllable canonical shape. Many other quite varied examples of this type can be shown for languages throughout the Austronesian family.

The disyllabic canonical target

In view of the disyllabic canonical target in Austronesian languages, the words that represent certain meanings are often conspicuous for their length. An example is the word for 'butterfly': Paiwan (Taiwan) *quLipepe,* Puyuma (Taiwan) *Halivanvan,* Bunun (Taiwan) *talikoan,* Ilokano (Philippines) *kulibangbang,* Tagalog (Philippines) *alibangbang,* Iban (Borneo and Malaysia) *kelebembang,* Tae' (Sulawesi) *kalubambang,* Sichule (Sumatra) *alifambang,* Gani (Halmahera) *kalibobo,* Numbami (north coast of New Guinea) *kaimbombo.* This word contains a prefix or family of prefixes that almost invariably is fossilized, thus creating a much longer word than is typical of Austronesian languages. The same phenomenon is seen with certain other meanings, such as 'ant,' 'firefly,' 'leech' (two types), 'echo,' 'dizzy,' and 'rainbow.'

In the Philippines clusters consisting of "heterorganic" consonants (consonants produced at different places in the mouth) are common in the middle of words (Tagalog *hagpós* 'loose, slack,' *puknát* 'unglued, detached'), but this is not typical of Austronesian languages in most other areas, where consonants tend to alternate with vowels.

Most Austronesian languages do not permit final palatal consonants, although in a few cases these have developed through secondary change. Other languages have a severely restricted inventory of possible final consonants in relation to consonants in other positions, as with Makasarese of southern Sulawesi, where the only possible final consonants are the velar nasal *-ŋ* and the glottal stop (a consonant produced by suddenly closing the vocal cords so as to interrupt the outward flow of air from the lungs).

In most Oceanic languages and some Austronesian languages in other areas all words end in a vowel. This is the result of either of two types of change: loss of final consonants or addition either of an "echo" vowel or of an invariant "supporting" vowel. Fijian and the Polynesian languages show open final syllables as a result of the first type of development; Mussau of western Melanesia and Malagasy show open final syllables as a result of the second type (see Table 56).

Table 56: Canonical Shape in Some Austronesian Languages

	Tagalog	Malay	Samoan	Mussau	Malagasy
'path/road'	da'ân	jalan	ala	salana	lalana
'sky'	lángit	langit	langi	—	lanitra
'salty'	ma-asín	masin	māi	masini	h-asina
'knock/beat'	tuktók	tutok	tutu	—	tutuka
'six'	aním	enam	ono	onomo	enina

PHONETICS AND PHONOLOGY

Size of phoneme inventory. Most Austronesian languages have between 16 and 22 consonants and 4 or 5 vowels. Exceptionally large consonant inventories are found in the languages of the Loyalty Islands in southern Melanesia, and exceptionally small consonant inventories in the Polynesian languages. Hawaiian has the second smallest inventory of phonemes, or distinctive sounds, of any known language, with just eight consonants (*p, k, ʻ* [glottal stop], *m, n, l, h,* and *w*) and five vowels (*a, e, i, o,* and *u*).

The second smallest inventory of phonemes

Vowel systems in Austronesian languages tend to be simple. Many languages in Taiwan, the Philippines, and Indonesia have just four contrasting vowels: *i, u, a,* and *e,* an indistinct mid-central vowel. The great majority of Oceanic languages have a five-vowel system: *i, u, e, o,* and *a.* Larger vowel systems are found in a number of Nuclear Micronesian languages, in some of the languages of Melanesia (such as Sakao of north-central Vanuatu), and in a few of the Chamic languages.

Phonetic types. In view of the large number of Austronesian languages it is not surprising that observers have recorded a wide range of speech sounds, including some that are quite rare in the world's languages. Most of the Formosan languages have a pharyngeal stop (written *q*), which is a consonant sound produced by drawing the backmost part of the tongue down to touch the wall of the pharynx. A number of the languages of Borneo and some other areas have unusual nasal consonants belonging to either of two types: "preploded" nasals, in which nasal consonants are heard as /-pm/, /-tn/, and /-kng/ at the end of a word, and what might be called "postploded" nasals /-mb-/, /-nd-/, or /-ngg-/, in which a nasal consonant between vowels is followed by a stop that is almost too short to hear.

Preglottalized or implosive consonants are found in several of the languages of central Taiwan, in a number of the languages of northwestern Borneo, in the Chamic languages of mainland Southeast Asia, and in several languages of the Lesser Sunda Islands. In Fijian and many other languages of Melanesia, voiced stops *b, d,* and *g* are automatically preceded by a nasal: *mb, nd,* and *ngg.* Perhaps the most unusual consonant types reported in Austronesian are prenasalized bilabial trills, made by trilling the lips following an *m,* and apico-labial stops (nasals and fricatives), which are made by touching the upper lip with the tip of the tongue. The former are quite common in the languages of Manus Island in the Admiralty Islands of western Melanesia, and the latter are found in a number of languages scattered throughout central Vanuatu.

Many Austroasiatic languages of the Mon-Khmer family found on mainland Southeast Asia distinguish two voice registers, a breathy, or "sepulchral," voice (made by relaxing the vocal cords) and a clear voice (made by tensing the vocal cords). As a result of generations of bilingualism this feature has been acquired by most of the Chamic languages. Together with other Mon-Khmer characteristics, these areal adaptations in the Chamic languages caused Schmidt in 1906 to incorrectly classify them as "Austroasiatic mixed languages." Where they have been further exposed to languages with lexical tone, as Eastern Cham (in contact with Vietnamese) or Tsat (in contact with both Chinese and Tai-Kadai tone languages on Hainan Island in southern China), at least two Chamic languages have become largely monosyllabic and tonal. Tonal contrasts are also reported for a few Austronesian languages in two widely separated parts of New Guinea and in southern New Caledonia. Despite contact with Chinese, which in some cases must date back at least three centuries, none of the aboriginal languages of Taiwan are tonal.

Contact with lexical tone

Many languages in the Philippines use stress to distinguish words that are otherwise identical in form, as in Tagalog *sâbat* 'design woven into cloth or matting' versus *sabát* 'stop pin or lug.' Some languages outside the Philippines use accent contrasts to distinguish different forms of the same word, as in Toba Batak (northern Sumatra) *gógo* 'push hard!' versus *gogó* 'strong' or *díla* 'tongue' versus *dilá* 'a big talker.' The origin and history of accent contrasts remains one of the major unresolved problems in the study of the Austronesian languages.

LEXICAL SEMANTICS AND SOCIOLINGUISTICS

Lexical semantics. Many common words in Austronesian languages are not easily translated into English or most other European languages. Examples of noncorrespondence can be seen in the comparison of several Malay words to English meanings: (1) one to many: Malay *kaki* corresponds to both 'foot' and 'leg' in English, (2) many to one: Malay *rambut* and *bulu* both correspond to English 'hair,' the former referring exclusively to hair of the head and the latter to body hair, downy feathers, plant floss, and the like, and (3) some combination of many to one and one to many: Malay *adik* corresponds to both 'brother' and 'sister' in English but is used only to refer to siblings younger than the speaker; Malay *kakak* also means both 'brother' and 'sister' but is used to refer to older siblings. In many Austronesian languages there is no general term for the verbs 'to cut' or 'to carry,' or for the noun 'root,' but rather numerous terms to specify the type of activity or type of structure in much greater detail than is typical in European languages.

Speech levels and honorific registers. Javanese and several languages in close contact with it—including at least Sundanese and Balinese—have developed a linguistic reflection of social stratification. Javanese uses three speech levels, distinguished by choice of vocabulary. The primary distinction is between Kromo, a high form used when speaking to social superiors, and Ngoko, a low or neutral form used when speaking to social equals or inferiors. Further subdivisions are recognized within Kromo, and in addition a small number of words called Madya (Middle) contain elements of both Kromo and Ngoko styles. In Samoa a special vocabulary is used when addressing persons of chiefly rank.

Male-female speech differences are covert in many languages, evident chiefly in the greater frequency with which speakers of one sex use particular forms; in some languages, however, gender-associated differences become conventionalized and rigid. The most notable case reported for an Austronesian language is in the Mayrinax dialect of Atayal in northern Taiwan, where women's speech is

Male-female speech differences

historically a more conservative variety, and men's speech shows unpredictable changes in pronunciation owing to the addition of entire syllables to earlier word forms.

These innovations present in Atayal men's speech may have originated as a form of speech disguise. In Tagalog and some other languages of the Philippines, as well as in Malay, forms of "backward speech" (which have as their primary purpose the concealment of messages) have been reported for adolescents. Such phenomena are functionally not unlike English pig Latin. Iban of northwestern Borneo shows an unusually large number of words with what appear to be reversals of the meanings found in cognates in other languages. This, too, may reflect an earlier tradition of speech disguise that succeeded in altering some meanings of the language for all speakers.

Reconstruction and change

GRAMMAR

Proto-Austronesian (PAN) probably had a verb–object–subject (VOS) word order. Four PAN affixes are commonly recognized: **Si-* marked instrumental focus (abbreviated IF), **-um-* actor focus (AF), **-an* locative focus (LF), and **-en* patient focus (PF). In addition, the infix **-in-* marked completive (c) aspect or past tense. The completive infix could co-occur with **Si-*, **-um-*, and **-an*, but, in the completive form of the patient focus, **-in-* was used without the suffix **-en*, and **-in-* thus simultaneously marked two functions: **k-um-aen i aku* (AF) 'I am eating,' **k-um-in-aen i aku* 'I was eating,' **kaen-en ni aku* (PF) 'is eaten by me/what I am eating,' **k-in-aen ni aku* (PFc) 'was eaten by me/what I ate.' This fusion of functions in the infix **-in-*, when used with the patient focus, has been carried down to many attested languages, including languages that no longer have a focus system.

Most views of grammatical change in Austronesian assume that Philippine-type focus systems continue a type of structure that was present from the earliest recoverable period. Not only do widely scattered languages, including Malagasy and Chamorro, have such systems, but many other languages have what appear to be fragments of a formerly more fully integrated system of particles and affixes. For example, in Tagalog the particle *si*, indicating actor focus for personal nominals, is syntactically opposed to *ni*, marking genitive/agentive. In Malay, a nonfocus language with a simple active-passive verb contrast corresponding to the focus systems of Philippine languages, *ni* has disappeared and the particle *si* has no grammatical function other than simply marking personal names or attributes used as names with a mildly pejorative connotation, as in *si Ahmad* 'Ahmad' or *si Gemuk* 'Chubby' (compare *gemuk* 'obese').

The functions of focus affixes

It is generally agreed that the focus affixes (with the possible exception of **-um-*) had both verbalizing and nominalizing functions. A more extreme view, not widely held, maintains that the focus affixes were originally used only to create nominals and were reinterpreted as verbal affixes in the separate histories of many daughter languages.

Proto-Oceanic diverged widely from this type of syntax. It appears to have been SVO, and most of the focus morphology of Proto-Austronesian was either lost or reinterpreted as nominalizing morphology. A major debate that has continued for three decades concerns the classification of various of the Polynesian languages as either accusative (having both transitive and intransitive subjects distinguished from objects) or ergative (having intransitive subjects and objects distinguished from transitive subjects).

MORPHOLOGY

The morphology of verbal focus has attracted the most attention in Austronesian studies, but other areas of morphology are also of interest. One such area is that of Ca-reduplication, a pattern of derivation in which the first consonant and vowel (stereotypically an **a*) are repeated. This pattern was first recognized with the numbers, where **esa* 'one,' **duSa* 'two,' **telu* 'three,' **Sepat* 'four,' **lima* 'five,' and the like are matched by a corresponding set of numbers **a-esa, *da-duSa, *ta-telu, *Sa-Sepat, *la-lima.* The unreduplicated set was used in serial counting or in counting nonhuman objects, and the reduplicated set in counting human beings. In some daughter languages (such as Tagalog) elements from both sets have survived and have been combined into a single set.

PHONOLOGY

Proto-Austronesian probably had the following consonant inventory: voiceless stops **p, *t, *C, *c, *k,* and **q;* voiced stops **b, *d, *z, *j,* and **g;* nasals **m, *n, *ñ,* and **ŋ;* fricatives **s, *S,* and **h;* liquids **l, *N, *r,* and **R;* and semivowels **w* and **y*. **C* and **c* probably were alveolar and palatal affricates; **q* was a pharyngeal stop. The **z* was most likely the voiced counterpart of **c*, while **j* appears to have been a voiced palatalized velar stop, a segment without any counterpart elsewhere in the system. The **s* probably was a palatal and **S* an alveolar sibilant; although conventionally written with the symbol for a nasal, **N* is more likely to have been a liquid of some kind; **r* seems to have been an alveolar tap, and **R* an alveolar or uvular trill. There were just four vowels: **i, *u, *a,* and **ə* (the schwa, a neutral mid-central vowel). In addition the semivowels **w* and **y* combined with **a, *i,* and **u* to form diphthongs **-aw, *-ay, *-iw,* and **-uy.*

From Proto-Austronesian to Proto-Malayo-Polynesian

The principal changes from this system to that of Proto-Malayo-Polynesian (the hypothetical ancestor of all non-Formosan Austronesian languages) are the merger of **C* and **t* as PMP **t*, the merger of **N* and **n* as PMP **n*, and the shift of **S* to PMP **h* (and of **eS* to **ah*). A number of other mergers occurred in Proto-Oceanic, including the merger of **p* and **b* and of **k* and **g;* the merger of the palatals **s, *c, *z,* and (in all Oceanic languages outside the Admiralty Islands of western Melanesia) **j;* and the merger of **e* and **-aw* as Proto-Oceanic **o.* These changes are illustrated in Table 57.

Table 57: Phonological Changes from Proto-Austronesian to Daughter Protolanguages†

	PAN	PMP	POC
'eye'	*maCa	*mata	*mata
'stone'	*batu	*batu	*patu
'ten'	*puluq	*puluq	*puluq
'rain'	*quzaN	*quzan	*qusan
'path/road'	*zalan	*zalan	*salan
'breast'	*susu	*susu	*susu
'crack open'	*pecaq	*pecaq	*posaq
'how many?'	*pija	*pija	*pija
'clothes louse'	*tumeS	*tumah	*tuma
'six'	*enem	*enem	*onom
'horsefly'	*laŋaw	*laŋaw	*laŋo

†PAN = Proto-Austronesian, PMP = Proto-Malayo-Polynesian, POC = Proto-Oceanic. An asterisk indicates a word that is reconstructed, not attested. A capital C indicates a sound similar to /ts/, a capital N probably was a type of /l/ distinct from the sound written as *l*, and a capital S indicates an alveolar sibilant similar in sound to English /s/.

VOCABULARY

About 5,000 unaffixed stems have been reconstructed for Proto-Austronesian, Proto-Malayo-Polynesian, or Proto-Western-Malayo-Polynesian. Although the Indo-European languages have a far richer textual tradition, probably no language family excels Austronesian in the richness of vocabulary reconstructed through the comparative method.

Reconstructed vocabulary shows clearly that the speakers of Proto-Austronesian had grain crops, including rice and millet; that they lived in settled villages of houses raised on piles; that they practiced weaving on simple back looms; that they domesticated dogs, pigs, and probably chickens; and that they were in contact with the sea and its resources. Familiarity with many tropical food plants can be inferred for speakers of Proto-Malayo-Polynesian. These include the coconut, banana, yam, sugarcane, pandanus, taro, sago, and breadfruit. Of these only sugarcane, pandanus, and wild taros of the genus *Alocasia* can safely be inferred for Proto-Austronesian, which probably was spoken on both sides of the Tropic of Cancer.

Reconstructions for 'boat,' 'sail,' and 'paddle' can be attributed to Proto-Austronesian, but terminology specific to the outrigger can be assigned only to Proto-Malayo-Polynesian, a language that was probably spoken somewhere

Lexico-statistical theory

in the northern Philippines in the period 3500–3000 BC.

Lexicostatistics, a controversial method for studying word replacement in relation to subgrouping, often distinguishes a subset of terms called "basic vocabulary." Lists of basic vocabulary words typically include those for body parts, terms for everyday natural phenomena (sky, wind, rain, sun, star, earth, stone, water, tree), basic kin terms (father, mother, child), and some others. Although lexicostatistical theory assumes a universally constant rate for the replacement of basic vocabulary, replacement rates in Austronesian languages appear to show considerable variation. Malay and its closest relatives (Iban, Minangkabau, and so on), many Philippine languages, and some languages in eastern Indonesia (Manggarai of the Lesser Sundas, Yamdena of the southern Moluccas) show very high concentrations of vocabulary items that have a wide distribution in the Austronesian family. It is inferred from this that they have replaced basic vocabulary at a slower rate than other languages. By contrast, languages in the South Halmahera–West New Guinea group and many of the Austronesian languages of western Melanesia show far lower concentrations of widely distributed forms, and it is inferred that they have experienced more rapid rates of basic vocabulary replacement. Some Oceanic languages—including several in the southeastern Solomons, Fijian, Polynesian (especially Samoan and Tongan), and the Chuukic (Trukic) languages of Micronesia—also have relatively large concentrations of widely distributed forms and have for this reason traditionally been highly valued as witnesses in comparative linguistics. (R.A.Bl.)

PAPUAN LANGUAGES

The Papuan languages are those languages spoken in an area centred upon New Guinea and extending from the islands of Alor, Halmahera, and Timor in the west to the Santa Cruz Archipelago in the east. The group includes approximately 740 languages, used by about 3,000,000 speakers. The term Papuan was originally employed merely to distinguish these languages from the Austronesian (Malayo-Polynesian) and Australian languages, and, until the second half of the 20th century, most Papuan languages were believed to be unrelated to each other. Intensive research by teams of linguists since the late 1950s, however, has resulted in a revolutionary change in the Papuan linguistic picture, and it is now known that about 350, and perhaps even as many as 450, of the approximately 740 identified Papuan languages are related. More than 500 of these are spoken by 2,900,000 speakers, who occupy almost three-quarters of the New Guinea mainland. Belonging to 76–105 language families (depending on the classification used), these languages together form the Central New Guinea macrophylum. (A macrophylum is a group of languages related less closely than those of a language family or stock.)

Recent extensive research into the Papuan languages has resulted in the preliminary classification of most of them. Numerous new languages have been discovered, though large areas, mainly in Irian Jaya, Indonesia, remain unknown. Despite the concentration of research on discovery and classification, a number of grammatical and lexical studies have been prepared, and folklore has been collected. The Summer Institute of Linguistics, an association of Protestant missionaries specializing in studying primitive languages and involved in literacy training and Bible translation, has carried out an extensive native language literacy program in New Guinea with a measure of success.

Classification and distribution. Interrelated language families are grouped in phyla, groups of languages more distantly related than those of a family or stock but more certainly or closely related than those of a macrophylum. Apart from the isolates, or unrelated languages, the number of known Papuan phyla is 21. Of these, 8 are small and consist of only 2 to 6 languages each. Of the remaining 13, 6 constitute the central New Guinea macrophylum, and 3 more can tentatively be included in it.

Central New Guinea macrophylum

The central New Guinea macrophylum comprises: (1) the East New Guinea Highlands phylum, found predominantly in the Highlands and Chimbu provinces of Papua New Guinea, (2) the Finisterre-Huon phylum in the Morobe and Madang provinces, Papua New Guinea, (3) the Central and South New Guinea phylum, located mainly in the Gulf and Western provinces of Papua New Guinea and in southern and northeastern Irian Jaya, (4) the West New Guinea Highlands phylum of the highlands area of Irian Jaya, (5) the South-East New Guinea phylum, in the Central, Northern, Morobe, and Milne Bay provinces of Papua New Guinea, and (6) the Madang phylum in the Madang provinces of Papua New Guinea.

Tentative member phyla of the Central New Guinea macrophylum are: (1) the Adelbert Range phylum in the Madang province of Papua New Guinea, (2) the Middle Sepik–Upper Sepik–Sepik Hill phylum, which could also be classified as constituting three separate phyla, and (3) the Anga stock (stocks are intermediate between phyla and families) spoken in the Eastern Highlands, Morobe, and Gulf provinces of Papua New Guinea.

Large Papuan phyla that are not members of the Central New Guinea macrophylum are the Ramu phylum in the Madang and East Sepik provinces of Papua New Guinea; the Torricelli phylum in the same provinces; the West Papuan phylum on the Doberai (Vogelkop) Peninsula of Irian Jaya and in northern Halmahera, eastern Timor, and Alor; and the Bougainville phylum on the island of Bougainville. Small isolated phyla are located mainly in the West Sepik and Gulf provinces of Papua New Guinea and in the Santa Cruz Archipelago. The highest concentration of known isolates is also in these provinces and in the East Sepik Province, the New Britain–New Ireland area, and the Solomon Islands.

Some Papuan languages are difficult to classify because of strong Austronesian influence upon them. Most Papuan languages are of only regional importance, but a few have achieved some cultural significance outside their immediate area because of their use as missionary languages.

Of the known Papuan languages that cannot yet be linked with the macrophylum, about 130 belong to large groups, and 26 to small individual groups. Approximately 50 languages, called isolates, seem to be unrelated either to each other or to the established groups. Further research may well show that additional Papuan languages belong to the macrophylum, especially after hitherto linguistically unknown areas in New Guinea have been studied.

Enga, the numerically largest language

Each of the individual Papuan languages is, for the most part, spoken by only a few hundred to a few thousand people, though the numerically largest one, Enga, in Enga Province of Papua New Guinea, has more than 150,000 speakers, and several other languages are each spoken by tens of thousands. Even if related, the languages generally show considerable diversity, especially in vocabulary. A few basic grammatical characteristics are, however, shared by many languages. In many instances it is difficult to determine the borderline between a language and a dialect, despite the presence of marked differences between them.

Linguistic characteristics. Most Papuan languages show extreme grammatical complexity. Their verbs vary to reflect a wide range of numbers and other features of the subject as well as of the direct and indirect objects and the beneficiary. For example, in Kiwai, a language of the Central and South New Guinea phylum, the verb *ai-ni-mi-bi-du-mo-iauri-ama-ri-go* means "they three will certainly see us two." Similarly, in the Monumbo language of the Torricelli phylum *mbepe*$_1$*-nge*$_2$ *tsi*$_3$*-p*$_4$*-ing*$_5$*-em*$_6$ can be translated as "you gave him a taro." Literally, the parts of this utterance are: taro$_1$–singular$_2$, 2nd person singular subject (you)$_3$–give, past form$_4$–masculine class 3rd person singular indirect object (to him)$_5$–plant class singular direct object$_6$. Verbs also indicate tense, aspect, mood, and the direction and circumstances in which the action they designate is performed; *e.g.*, in Gadsup, of the East New Guinea Highlands phylum, *kùmù-ə̃-nk-ə̃də̃d-òn-ték-ə̃p-*

ón-ì-nó-ké, "had he indeed wanted to go down for him?", has several prefixes and suffixes indicating emphasis, tense, direction, and the question status. (The accent marks indicate tones; the ə is a sound pronounced like *a* in "sofa." The dashes separate the various components of the utterances, but do not normally appear in nonlinguistic texts).

Verb forms for one or more actions

There are, basically, two major types of verb forms in many Papuan languages. One, which can be referred to as the normal type, is found in sentences in which only one action is referred to; the other, which may be referred to as the special verb form, occurs in sentences in which more than one action is mentioned. In the latter type of sentence, which may have one or more special verb forms, the normal verb form also appears with the last verb in the sentence.

Numerous Papuan languages have gender and noun class systems, some with up to ten or more classes. Bewilderingly complex variations of adjectives, numerals, demonstratives, and subject and object markers often result, since these words have special forms for each of the various classes of nouns. An example is the sentence *ame akwum kuvambakwum sumupar amenakwum salikəmba,* "I saw my two big women," from Angoram of one of the small Sepik phyla. If "women," *akwum,* is replaced by "arrows," "gardens," or "frogs," etc., the entire sentence, rather than a single word, is changed: when "arrows" is substituted, the phrase becomes *ame pwanggli kəpanggli klupar amenakanggli salikənggliya;* and when "gardens" is used, it changes to *ame konggəmbər kəvambər pələpar amenkəmbər salikəmbəra.*

Many languages are tonal, with changes of pitch in words and syllables that affect the meaning of the words. The interaction of Papuan tonal systems with patterns of stress and syllable length can be extremely intricate.

AUSTRALIAN ABORIGINAL LANGUAGES

There are approximately 260 Australian Aboriginal languages. This group of genetically interrelated tongues embraces the entire Australian continent as well as the western islands of Torres Strait, but apparently excludes Tasmania. The languages are characterized by great similarities in their sound systems and considerable agreement in grammar but often by markedly few similarities in vocabulary. Intelligibility between neighbouring forms of speech is common, and dialect chains stretching over amazing distances occur, though the two extremes of such a chain seem to be quite distinct languages.

Secret languages and special vocabularies

Every tribe speaks at least one distinct dialect, but bilingualism and multilingualism are common in many areas. Many individual languages have parallel forms, characterized by special vocabularies and sometimes by special sounds that are used in cultural avoidance situations (*e.g.,* to mothers-in-law) or as secret languages among initiated men on certain occasions.

No genetic link is known to exist between the Australian languages and any outside language. It is believed that languages ancestral to the present-day ones were introduced into Australia by peoples that crossed Arnhem Land in northern Australia many millennia ago. With the apparent exception of the influence of Papuan languages on the languages of the Cape York Peninsula, the Australian languages remained free from outside influence until the arrival of European settlers late in the 18th century.

The great majority of the Australian languages were nearing extinction by the third quarter of the 20th century, with about 50 or more extinct, predominantly in the east, south, and west of the continent. Speakers of languages believed extinct for decades are, however, occasionally discovered. Most languages have very few surviving speakers; still-vigorous languages have, for the most part, only a few hundred speakers each, though Mabuiag (the language of the western Torres Strait islands) and the Western Desert language have 8,000 and 4,000 speakers, respectively. About 45,000 Aborigines may still have some knowledge of an Australian language, but accurate figures of the speakers of individual languages are almost impossible to obtain.

Extensive research on the Australian languages has been carried out since 1960, largely through the Australian Institute of Aboriginal Studies in Canberra. The results of this and earlier research have shown the Australian languages to be interrelated and have made it possible to explain their structural differences in terms of a typological development from a simple to a complex structure. In addition, a considerable amount of detailed information on the grammar of numerous Australian languages has been recorded.

Classification. Earlier classifications regarded the northern and northwestern languages, which are structurally rather different from the southern languages, as genetically distinct from the latter. (The interrelationship of the southern languages was recognized quite early.) Later, various classifications based on language type were established, and these demonstrated the ultimate unity of all Australian languages. The most recent classification is based on the degree of lexical (vocabulary) interrelationship between the languages; it subdivides the languages into 28 families, of which 27 are located in the north and northwest, covering about one-eighth of the continent, and a single family is found occupying the remaining seven-eighths of Australia. This skewed picture may be the result of the thorough spreading of a language form referred to as Common Australian (dated at about 5,000–6,000 years ago) from somewhere in northwestern Australia through most of the continent except the north and northwest regions. This diffusion appears to have coincided with that of an archaeologically recognized cultural revolution. The spreading of this Common Australian language may have brought about greater linguistic uniformity in much of Australia. Most of the 28 families are subdividable into groups and often into subgroups of individual languages. In the following list of language families these abbreviations are used: G = group or groups; SG = subgroups; L = language or languages.

Genetic relationship of all Australian languages

Tiwian (1L)	Djingili-Wambayan (2G, 3L)
Yiwadjan (4G, 2SG, 5L)	Karawan (2G, 2L)
Kakadjuan (1L)	Minkinan (1L)
Mangerian (2G, 2L)	Larakian (2G, 2L)
Gunavidjian (1L)	Kungarakanyan (1L)
Nagaran (1L)	Warraian (1L)
Gunwingguan (6G, 3SG, 11L)	Daly (3G, 4SG, 10L)
Bureran (2G, 2L)	Murinbatan (1L)
Nunggubuyuan (1L)	Djamindjungan (4L)
Andilyaugwan (1L)	Djeragan (2G, 5L)
Maran (2G, 2SG, 3L)	Bunaban (2G, 2L)
Mangaraian (1L)	Wororan (3G, 12L)
Ngewinan (1L)	Nyul-Nyulan (4L)
Yanyulan (1L)	Pama-Nyungan (41G, 50SG, 177L)

Some aspects of this classification are only tentative. The locations of the families are shown on the accompanying map.

In most areas in which Australian languages are still in daily use, individual languages have gained prominence over others and have become lingua francas (common languages), as a result either of their use as mission languages or of social and cultural factors active among the Aborigines themselves. One of the dialects of the Western Desert language, Bidjandjara (Pitjantjatjara), has become the Aboriginal lingua franca over a sizable portion of the western half of the continent.

Bidjandjara

Grammar. The Australian languages generally show considerable grammatical complexity. Affixes (word parts added initially, internally, or terminally) play an important role; prefixes used initially and suffixes used terminally are found in northern and northwestern Australia, and suffixes, for the most part, are used elsewhere. A peculiar feature of many languages is the suffixing of markers indicating the subject and object of the verb to

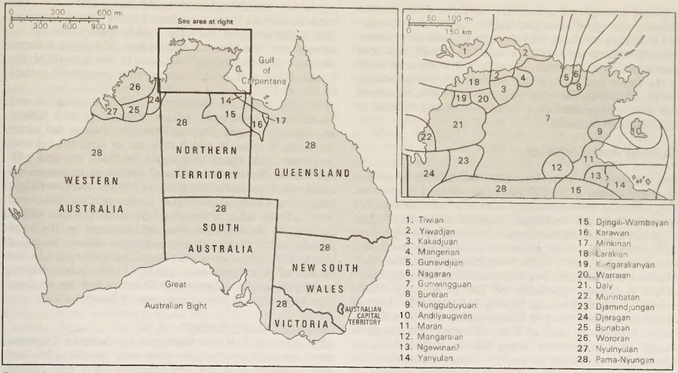


Figure 29: Distribution of the Australian Aboriginal languages.
Adapted from S.A. Warm, *Languages of Australia and Tasmania* (1971); Mounton & Co.

the first word of the sentence, irrespective of what it is, or to special particles not connected with the verb. An example from Wanman of the Pama-Nyungan family is the use of the suffixes -ŋ*a* and -ŋ*ku* in *paral*^{y}*i*-ŋ*a*-ŋ*ku* *t*^{y}*i*ŋ*ka*-ŋ*a*, literally, "boomerang-I-you make-past," or "I made a boomerang for you." Another widespread feature is an ergative or agentive suffix, attached to nouns and pronouns, that indicates the actor of an action referred to by a transitive verb. In the Dungidjau language of the Pama-Nyungan family, *t*^{y}*a:n-tu pukin*^{y}*-n*^{y}*a pumi* is directly translated as "*man*-ergative *dog*-object *hit*-past." Here *-tu* is the ergative suffix; attached to *t*^{y}*a:n* ("man"), *-tu* signifies that "man" is the actor of the verb, thus rendering the meaning "the man hit the dog." Great freedom of word order is a feature of many Australian languages. A number of languages, mainly northern ones, have gender and noun class systems, with adjectives, numerals, and demonstratives showing special forms for each of the classes of nouns and often, also, for their number. For example, in Andilyaugwa of the Andilyaugwan family, "where is (located)" is expressed as ŋ*i*-ŋ*ampa na-mpil*^{y}*a* when referring to one man, but as *wunala*-ŋ*ampa wu-pil*^{y}*a* when referring to two men, *wura*-ŋ*ampa na-mpil*^{y}*a* in regard to three or more men, *ta*-ŋ*ampa i*ŋ*a mpil*^{ys}*a* to one woman, and *ma*-ŋ*ampa numa-mpil*^{y}*a* to a ship.

Phonology. The sound systems of the Australian languages are extremely similar. Most of them share from four to six different points of articulation for stop consonants (made with complete stoppage of the breath from the lungs) and nasal consonants (made with the airstream passing through the nose), with many languages having such consonants produced with the tip of the tongue placed between the teeth (interdental consonants), or behind the teeth (alveolar consonants), indicated as *t, n,* or curled up against the hard palate (retroflexed consonants), written as *ṭ, ṇ*. The series of consonants may thus include labial consonants (*p, m*), interdental consonants, alveolar consonants (*t, n*), retroflexed consonants, palatalized consonants (*t*^{y}, *n*^{y}), and velars (*k*, and ŋ as the *ng* in "sing"). In addition, most Australian languages show no distinction between voiced and voiceless stops (such as voiced *b* and voiceless *p* in English), no fricative consonants (*e.g., f, v, s, z*), only three vowels (*a, i, u*), but two or three distinct *r* sounds.

Common Australian words

Vocabulary. In spite of the great vocabulary differences among the Australian languages, a number of common words are encountered in a great many languages all over the continent. These are believed to constitute a Common Australian element. In the vocabulary of a given language, various classes of words, such as nouns, verbs, and others, are clearly distinguishable and definable, and word formation is through the use of affixes. Australian languages have contributed to Australian English mainly animal and plant names and objects in nature—kangaroo, wallaby, kookaburra, budgerigar, galah, coolibah tree, billabong.

Collections of mythological and other text materials have been and still are being made in a number of languages. Native literacy in Australian languages is limited but is on the increase as a result of the work of members of the Summer Institute of Linguistics (an association of Protestant missionaries that specializes in studying primitive languages), who, like the people from the Methodist, Anglican, and Australian Inland missions, write the languages in specially adapted versions of the English alphabet.

(S.A.W.)

AFRICAN LANGUAGES

Although no definitive count exists, the number of separate and distinct African languages, as estimated by various authorities, ranges from 800 to more than 1,000. The linguistically most homogeneous area of Africa is the northern part, in which Arabic predominates from Egypt to Mauritania, albeit in a number of sharply distinguished dialects; the most important dialect cleavage is between the Egyptian–Sudanese varieties and those of the Maghrib (from Libya westward). Intermingled with the Maghrib dialects are the Berber languages, concentrated principally in Algeria and Morocco. They range as far east as the Siwa Oasis in western Egypt, as far west as the Senegalese–Mauritanian border area, and as far south as the southern rim of the Sahara. The nomadic Tuaregs speak a Berber language. Nubian, a totally different language, is spoken as far north as southern Egypt along the Nile, but its links are clearly with the languages of sub-Saharan Africa.

Sub-Saharan languages

In sub-Saharan Africa the linguistic picture is far more complex. Except for the Khoisan (Bushman and Hottentot) languages of the extreme south, approximately the entire southern third of Africa is occupied by the relatively closely interrelated Bantu languages. They extend eastward from roughly the Nigerian–Cameroon border to the Indian Ocean. Bantu and non-Bantu languages are considerably interspersed around the northern Bantu border, running through Cameroon, slightly south of the northern boundary of Zaire, and through Uganda and Kenya. There are Bantu enclaves as far north as Somalia and non-Bantu enclaves in northern Tanzania. It is the area north of the Bantu and south of the Sahara that is linguistically the most diverse; there the languages are most numerous and their interrelationships most remote and difficult to establish with certainty.

General considerations

African lingua francas

Because probably not more than 40 African languages have more than 2,000,000 native speakers, a number of lingua francas have developed in various areas as a means of coping with this enormous linguistic diversity. Arabic, in addition to having the largest number of native speakers of any language of the continent, is sometimes used as a lingua franca, in its literary form by educated non-Arab Muslims, and in a number of colloquial varieties by much of the non-Arab population of The Sudan and Chad. Swahili, a Bantu language heavily influenced by Arabic, is official in Tanzania and Kenya and is used as a lingua franca throughout most of East Africa including eastern Congo. Other important lingua francas include Lingala, also a Bantu language, in western Congo; Fanagalo (vulgarly known as "kitchen Kaffir" or "mine Kaffir"), a pidginized form of Zulu with many English and Afrikaans loanwords, in South Africa, particularly in the mines; and Sango, a pidginized form of Ngbandi (included in the Adamawa-Eastern subdivision of Niger-Congo by some scholars; see below) with many French loanwords, spoken in the Central African Republic.

In addition, there are Bambara-Maninka, a Mande language used in Mali, Guinea, and Ivory Coast; Hausa, spoken in northern Nigeria and neighbouring areas; Amharic, used throughout Ethiopia, where it is the official language; Wolof, the language of Dakar, widely spoken throughout Senegal and The Gambia; and Kongo (Kikongo), a Bantu language that is used in the area of the mouth of the Congo River and is also spoken in a pidginized variety known as Kituba.

Standard European languages, particularly English and French and, to a lesser extent, Portuguese and Italian, also serve as lingua francas, generally among the better educated, in areas in which they are or were the colonial languages and where they enjoy or have enjoyed official status. Localized varieties of European languages are also spoken in numerous parts of Africa, such as Pidgin English, a widely used lingua franca of Cameroon and West Africa; Krio, a creolized form of English, differing slightly from Pidgin English and used by the local population of Freetown, Sierra Leone; various forms of Creole Portuguese, spoken in the Cape Verde Islands, the islands of São Tomé and Príncipe, and Guinea-Bissau; and Afrikaans, an outgrowth of Dutch, which serves along with English as an official language of South Africa.

LANGUAGE CLASSIFICATION

Early attempts at classification

Although word lists of a number of sub-Saharan languages were recorded by Europeans as early as the 15th and 16th centuries, and the first grammar of a sub-Saharan language (Kongo) was published in 1659, extensive study of African languages and systematic attempts at grouping and classifying them did not begin until the 19th century. In 1854 Sigismund Koelle, a missionary living in Freetown, Sierra Leone, was able to collect substantial word lists of roughly 150 languages, principally of West Africa but also including some belonging to the Bantu family. He established 11 groups, leaving approximately 40 languages

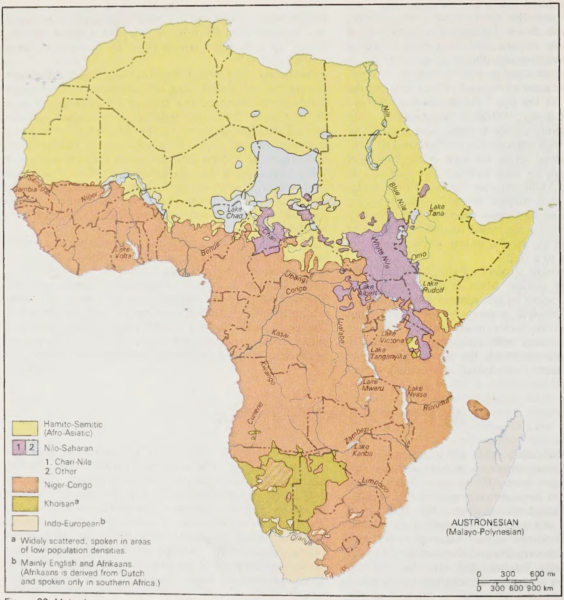


Figure 30: Major language groups in Africa.

unclassified. His groups are generally small, rarely exceeding 15 members, and he made no systematic attempt to explore wider relationships. He laid the groundwork for establishing the Mande family, which has been universally accepted ever since. (The term "family" is used loosely in this article, sometimes referring to large groupings of languages and sometimes referring to subgroups or branches of these groupings. Niger-Congo, for example, is usually called a language family, but its subgroups [*e.g.*, Mande] and its sub-subgroups [*e.g.*, Bantu] are often referred to as families as well.)

The unity of the Bantu family, although suggested by a number of earlier writers, was first established in 1862 by the German scholar Wilhelm Bleek, who gave it its name, the word for "people" in most languages of the family. During approximately the same period, the work of several scholars suggested that Ancient Egyptian, the Berber languages, certain languages of northeastern Africa called Cushitic (*e.g.*, Galla, Somali, Beja, and Afar [Danakil]), and perhaps others (*e.g.*, Hausa) were remotely related to one another and to the Semitic family. These languages were called Hamitic at the suggestion of a French Semitist, Ernest Renan, who was also the first to propose the name Cushitic, because the peoples speaking these languages coincided roughly with the descendants of Ham (one of whom was Cush), according to the biblical account in Genesis, just as the people speaking the Semitic languages coincided with the descendants of Shem (Sem).

The first attempts at an overall classification of African languages were those made by the Austrian linguist Friedrich Müller in 1876–88 and by the German Egyptologist Karl Richard Lepsius in 1880. Müller, who classified all the world's languages on the basis of a correlation of linguistic and racial groupings (the latter determined largely by hair type), set up the following divisions within Africa: (1) Bushman-Hottentot, as a subdivision of the languages of the tufted-haired peoples (the Papuan languages of New Guinea formed the other subdivision); (2) Bantu and (3) Negro, both subdivisions of the languages of the fleece-haired (*vliesshaarig*) peoples; (4) Hamitic, (5) Semitic, and (6) Nuba-Fulah, all three subdivisions of the languages of the wavy-haired (*lockenhaarig*) peoples. Linguistically, Nuba-Fulah was completely heterogeneous and was abandoned by later scholars.

Lepsius' classification

Lepsius proposed an alternative classification of the languages and peoples of Africa, of which the three major divisions were (1) Semitic; (2) Hamitic, including Hausa, which he grouped with the Berber or Libyan branch, and Hottentot, neither of which had been classed as Hamitic by Müller; and (3) the Negro languages, consisting of two subdivisions, Bantu and Mixed Negro, composed of many subgroups, and containing all the indigenous languages of Africa not included in the other divisions.

During the early part of the 20th century, a revision of this classification emerged from the work of the two leading German Africanists, Carl Meinhof and Dietrich Westermann. Meinhof added a number of languages to Lepsius' Hamitic grouping, namely Fulani, a major West African language, and a group of East African languages, including Masai and Bari. Westermann attempted to show the unity of virtually all the non-Bantu, non-Hamitic, and non-Semitic languages of Africa; he designated them Sudanic. Thus, a threefold division was established: Bantu, Hamitic, and Sudanic. The Bushman languages of southern Africa were sometimes treated as a separate group and sometimes linked with Sudanic, and Semitic was generally linked to Hamitic, though not included within it. Not everyone accepted this classification. French scholars, in particular Maurice Delafosse and Lilias Homburger, restricted Hamitic to Ancient Egyptian, the Berber languages of North Africa, and the more heterogeneous Cushitic group of languages of northeastern Africa. They connected Hottentot with the Bushman languages while grouping the other languages classed by Meinhof as Hamitic in the very general Negro-African or Sudano-Guinean family, to which they also believed the Bantu languages to be affiliated.

Westermann himself revised his views of the earlier classification scheme and came to regard the great majority of the languages of West Africa (*i.e.*, the Western Sudanic languages) as more closely related to Bantu than to the languages farther east, previously regarded as Sudanic. Sir Harry Johnston, a British Africanist, likewise pointed out similarities between Bantu and many West African languages that he termed Semi-Bantu.

Joseph Greenberg's classifications

No single scheme was generally accepted in 1955, when Joseph Greenberg, a United States linguist, presented the earlier version of his major classification of the languages of Africa. In 1963, Greenberg made public a substantially revised version of his work that consolidated previously unclassified languages and smaller language families into four large linguistic stocks—Niger-Kordofanian, Nilo-Saharan, Khoisan, and Afro-Asiatic (Hamito-Semitic). They included all the indigenous languages of Africa except Malagasy, the language of Madagascar, which belongs to the Austronesian (Malayo-Polynesian) family and is thus unrelated to the languages of the continent. The Niger-Kordofanian group consists of two branches, Niger-Congo, which is very extensive, and the much smaller Kordofanian group, so named because all of its members are located in Northern and Southern Kurdufān provinces, The Sudan. Kordofanian contains a number of subgroups, of which Kadugli-Korongo is most distantly related to the others. In fact, its inclusion within the group is open to serious question. Niger-Congo consists of Westermann's Western Sudanic group plus Bantu, whose interrelationship had earlier been advocated by Westermann. Greenberg differed from Westermann, however, in that he treated Bantu as a subgroup of one of the branches of Western Sudanic rather than an independent unit related to Western Sudanic as a whole. Greenberg also argued strongly for the inclusion of Fulani in this family. (Fulani had been regarded as Hamitic by Meinhof, but some French scholars had already advocated its relationship to other West African languages, such as Serer of Senegal.) Furthermore, Greenberg extended Niger-Congo eastward to include languages of Central Africa, such as Ngbandi of the Central African Republic and Azande (Zande) of Zaire, with which Westermann had not dealt.

The Nilo-Saharan group consists of six branches, of which Chari-Nile is by far the most heterogeneous. This family is the least adequately substantiated of the four. The Khoisan languages comprise the Bushman and Hottentot languages of southern Africa and two languages of Tanzania, Sandawe and Hadza (Hatsa). Afro-Asiatic, which is Greenberg's name for Hamito-Semitic, consists of five branches: (1) Semitic, represented in Africa by Arabic and the Ethiopic languages, of which the most important are Amharic, the official language of Ethiopia; Tigrinya, the predominant language of Eritrea; and Geʿez, the liturgical language of the Ethiopian Church, which is no longer spoken; (2) Ancient Egyptian, now extinct, and its daughter language, Coptic, the liturgical language of the Coptic Church of Egypt, (3) the Berber languages of northern Africa, including Kabyle of Algeria, Riffian and Shluh (Shilha) of Morocco, and Tuareg, or Tamashek, of the Sahara; (4) the extensive and diversified Cushitic group of northeastern Africa, whose most important members are Somali and Galla; (5) Chadic, spoken in Chad, Cameroon, and northern Nigeria, although Hausa, its best known and most widely spoken member, is used throughout much of West Africa.

Branches of Afro-Asiatic (Hamito-Semitic)

Hamito-Semitic languages also are spoken in the Middle East. For this reason, they are treated in a separate section (see *Hamito-Semitic languages*).

Because of the inexact nature of the procedures for establishing genetic classification and also because of incomplete data on the majority of African tongues, differences of opinion exist on various aspects of Greenberg's classification. Some scholars, such as the Hungarian linguist István Fodor and the English Bantuist Malcolm Guthrie, insist that genetic interrelationship of a group of languages can be established only by demonstrating regular phonological correspondences among them, as have been shown among the Bantu languages, for example. Greenberg and his supporters, however, maintain that similarities in both sound and meaning between two or more languages, which are too numerous to be dismissed as mere coincidence and are found in that part of language least susceptible to borrowing (*e.g.*, grammatical elements, body parts and functions, common natural objects and phenomena, and lower numerals), must be attributed to genetic relationship. Nevertheless, there are no rigorous procedures for determining how similar two items must be in sound and meaning to be considered evidence for genetic relationship or how many items are needed to rule out other explanations.

SIMILARITIES AND CORRELATIONS

Common features of the African languages. Although African languages are extremely diverse in structure, certain phonological and grammatical features are widespread throughout much of the continent or at least within extensive and well-defined areas.

Tones and clicks

Phonological features. Tone—*i.e.*, the use of pitch to distinguish words and grammatical forms that otherwise would be phonetically identical—occurs in the overwhelming majority of sub-Saharan languages. The famous click sounds are largely restricted to two language groups of southern Africa: the Bushman and Hottentot (Khoisan) languages and the southern Bantu languages (*e.g.*, Zulu, Xhosa, and Sotho), which borrowed them from the Bushman- and Hottentot-speaking peoples. Clicks are also common to three East African languages, Sandawe, Hadza, and Sanye, but nowhere else on the continent or outside it. The consonants *kp* and *gb*, called coarticulated labiovelar consonants, are, in Greenberg's words,

> distributed over a wide belt of languages of diverse genetic affiliations from the Atlantic Ocean almost to the Nile Valley. Outside of Africa these sounds are only known from a restricted area of Melanesia. . . .

Other phonological features common in Africa but rare in the rest of the world are word-initial combinations of a nasal sound plus another consonant (*e.g.*, *mb* or *nd*), and implosive consonants.

Grammatical features. In some African languages, words generally consist of a root only, with grammatical affixes rare and even nonexistent in a few languages. These are often referred to as isolating languages. Other African languages have words that are composed of many elements. Those languages are subdivided into the agglutinative languages (in which each element of a word has a distinct and separate meaning and form) and inflectional languages (in which the various elements of meaning may be fused into such forms as prefixes or suffixes or the entire word may change internally to indicate grammatical relationships).

In some of these languages, a single word might be equivalent to a whole sentence in English or other European languages. Thus, in Swahili, a Bantu language, "we will not hit him" is expressed in a single word, *hatutampiga*, in which *ha* is the negative marker, *tu* the verb subject "we," *ta* the future-tense marker, *m* the verb object "him," and *piga* the verb "hit." The verb likewise may undergo modifications in meaning through the addition of suffixes; thus *pigwa* means "(to) be hit," and *pigana* "(to) hit one another."

Noun classes and gender

A widespread feature among African languages employing words composed of more than one element is a system of noun classes (see below *Benue-Congo subgroup*), particularly marked in the Bantu family but also occurring in a wide variety of West African languages, as well as in certain languages of the Northern and Southern Kurdufān provinces, The Sudan. All of these languages have been grouped by Greenberg in his Niger-Kordofanian family, although others in this family lack this feature altogether.

Grammatical sex gender is not nearly as widespread as noun classes. It is characteristic of the Semitic and Hamitic languages (although lacking in some members of the Chadic branch) and of other languages such as Hottentot and the closely related Naron Bushman language of South Africa and of the much more distantly related Sandawe and Hadza languages of Tanzania. Gender is likewise found in East Nilotic but not South Nilotic (these two comprise what was earlier known as "Nilo-Hamitic") or West Nilotic (previously known simply as Nilotic). Gender plays an even more restricted role in other African languages, such as Bongo, a language of the southwestern Sudan in which different pronouns are used for masculine and feminine 3rd person singular. The great majority of African languages, however, do not make this distinction, but many of them do distinguish between animate and inanimate personal pronouns.

Linguistic, racial, and cultural interrelations. It is difficult to find strict correlations between linguistic, cultural, and racial groupings, either in Africa or elsewhere in the world. Groups speaking identical or closely related languages may be strikingly different racially, culturally, or both; while groups that are very similar racially and culturally may speak different and even unrelated languages. Thus the Damara (or Bergdama) of South West Africa/Namibia, who speak the language of the Nama Hottentot, are sharply distinguished from the latter both racially and culturally. The Damara have much darker pigmentation and other more Negroid physical features and are predominantly nonpastoralists. Likewise, the Arusha of northern Tanzania, who speak the language of the pastoralist Masai, are horticultural (*i.e.*, agricultural without use of the plow). On the other hand, the forest-dwelling Pygmies, who are spread over much of central Africa, form a distinct racial and cultural group and are exclusively hunters and food-gatherers, yet they have no distinctive language. Instead, they speak the languages of their various neighbours, such as Bantu, Moru-Mangbetu (grouped by Greenberg in the Central Sudanic branch of Chari-Nile), and reportedly Sere-Mundu (grouped by Greenberg in the Adamawa-Eastern branch of Niger-Congo).

Broad correlations between race, language, culture

Some broad correlations can nevertheless be made between language groupings and the variables of race and culture, as long as these are not interpreted too rigidly. The Bushmen of southern Africa form a distinct racial, cultural, and linguistic group, and, though the Hottentot are akin to them linguistically and racially, they show important cultural differences, most notably that the Hottentot are pastoralists, not hunters and gatherers. The Berber-speaking peoples of North Africa likewise form a relatively homogeneous racial, cultural, and linguistic group; they are Caucasoid and pursue a mixed agricultural and pastoral way of life. Only the Sahara-dwelling Tuareg are exclusively pastoral nomads and caravaners. The Cushitic peoples of northeastern Africa, who like the Berbers speak languages of the Hamito-Semitic family, represent varying degrees of Caucasoid and Negroid elements. Some of the more prominent among them, such as the Beja or Bedawiye of The Sudan and the Saho-Afar and Somali of the Horn, are nomadic pastoralists. The Galla of southern Ethiopia, the most numerous of all Cushitic peoples, include some pure pastoralists, but the great majority practice agriculture and husbandry. Their neighbours to the north, the Agau of central Ethiopia, do likewise, though they have largely assimilated culturally to their Semitic-speaking neighbours, while retaining their Cushitic form of speech.

The unusually tall and slender Nilotes of The Sudan, such as the Dinka and Nuer, are sometimes considered a subrace of the Negroid; but it has also been suggested that their physical peculiarities are at least to some extent the result of differences in diet rather than heredity. Herding of cattle and other livestock is predominant among the majority of Nilotic peoples (including the socalled "Nilo-Hamites"), but nearly all practice some horticulture, with the striking exception of certain groups of Masai. Certain Nilotic peoples, however, are predominantly sedentary horticulturalists, such as the Anuak of The Sudan and the Luo, principally in Kenya, who are the most numerous of the Nilotic peoples.

The overwhelming majority of the Bantu and the peoples of western and central Africa are sedentary horticulturalists. Likewise, most of these peoples belong linguistically to Greenberg's Niger-Congo family, although some fall within Nilo-Saharan, and others within the Chadic branch of Hamito-Semitic; furthermore, they comprise the overwhelming majority of the Negro race, which, however, is no more homogeneous a physical group than the peoples

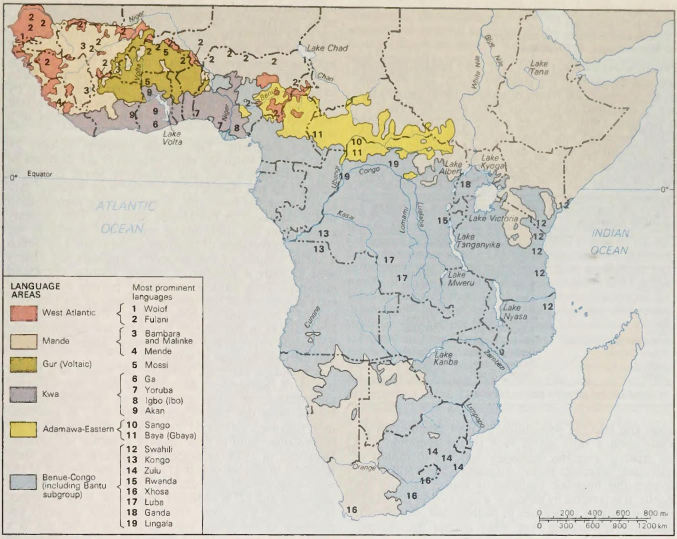


Figure 31: Distribution of the Niger-Congo languages.

West Africa are the cattle-herding Fulani, who are widely distributed over the savanna belt just south of the Sahara. They are physically as well as culturally distinct from their neighbours, but linguistically their closest affinities are with the sedentary horticultural peoples of Senegal. Cattle raising likewise plays an important role among certain Bantu peoples of eastern and southern Africa, such as the Ganda, Nkole (Ankole), Sotho, and Nguni peoples. Among some of the East African Bantu (*e.g.*, the Nkole, Rwanda, and Rundi) cattle ownership has been the prerogative of a privileged group probably of Nilotic origin and known by various names, such as Hima (Panda) and Tutsi; these people are culturally distinct from their agricultural neighbours, although totally assimilated to them linguistically. (M.F.Go.)

Niger-Congo languages

The Niger-Congo family is the largest family of languages in sub-Saharan Africa, with 890 known member languages. All of them are considered to be distinct languages and not simply dialects. The named dialects of these languages number in the many thousands; if the variant names for the same language or for the same dialect are to be counted, the figure must be further expanded. For example, for Swahili, there are some 15 separate dialects, additional variant names for some of the dialects, and at least four "nicknames" for popularized or debased forms of the language.

The classification of such an enormous and diverse family was first accomplished in 1949 by Joseph Greenberg, who demonstrated the genetic unity of this group on the basis of an examination of documentation of all the languages of Africa. The primary sources for such a massive comparison consist of published descriptions or vocabularies of tribal languages made by missionaries and colonial officers; these were written in European languages because practically none of the African languages had any written tradition of their own. Usable documents stretch back for only about 125 years, and there are still a number of indigenous languages that either have not been reported or lack sufficient detail or accessibility to be included in any final count of Niger-Congo languages. As a consequence, the figure given here may have to be revised upward, especially as there is no evidence that the indigenous languages of sub-Saharan Africa are dying out. Quite to the contrary, of a list of such languages collected during the mid-19th century, only one is known to have become extinct.

Location of the Niger-Congo languages

The area in which the Niger-Congo languages are spoken is fairly easily demarcated (see map) and includes the entire sweep of sub-Saharan Africa with the exception of the Horn of Ethiopia. All indigenous peoples of sub-Saharan Africa west of the Nile are speakers of Niger-Congo languages with the exception of a few to the extreme south (San and Khoisan peoples) and a more substantial number along the northern rim of the area (speakers of Chadic or Nilo-Saharan languages).

A conservative estimate of the total number of speakers of Niger-Congo languages is 300,000,000. Such a figure is obtained by adding the populations (as of the early 1980s) of the 32 countries of tropical and southern Africa and then subtracting the number of European and Asian speakers reported for South Africa. (It is assumed that the number of speakers of Niger-Congo languages in the countries to the north of the tropical belt omitted by this method—especially in Mali and Burkina Faso—would be balanced out by the number of indigenous non-Niger-Congo speakers enumerated in the included populations—especially in Nigeria and Uganda.)

CLASSIFICATION

The Niger-Congo family is divided into six genetic subgroups. Viewed on a geographic continuum from northwest to southeast, these are:

1. West Atlantic, 43 languages
2. Mande, 26 languages

3. Voltaic (Gur), 79 languages
4. Kwa, 73 languages
5. Adamawa-Eastern, 112 languages
6. Benue-Congo, 557 languages

This subclassification results from very fundamental and very old divisions in the language family, and each subgroup represents the results of separate developments that can be estimated to be about 5,000 to 10,000 years old at the least. Obviously, such a deep division has yielded widely separated languages—separate not only in geographic distribution but in elements of the sound systems, grammars, and vocabularies as well. It is surely safe to say that there is no one feature common to the entire group and that the recognition of their relationship depends upon a truly global view of the evolution of a few separate ancestral languages into the thousands of descendant languages spoken throughout the world today. Because this linguistic evolution is neither random nor chaotic, it is possible by means of a very careful plotting of common clusters of features to establish the approximate degrees of relationship even when neither the intermediate nor earliest stages of the protolanguages concerned have been reconstructed. Greenberg used this method to include the nearest relatives of the Niger-Congo languages with them in a superfamily that he designates as "Niger-Kordofanian." This grouping contains two branches: all the languages here considered, namely the Niger-Congo family, and a coordinate branch of the Kordofanian family of languages, spoken in the Nuba Hills (Jibāl an-Nūbah) of Southern Kurdufān province, The Sudan.

Table 58: Niger-Congo Languages with More Than Five Million Speakers

	subgroup	estimated number of speakers*
Fulani (Fulbe)	West Atlantic	11,500,000
Yoruba	Kwa	18,100,000
Igbo (Ibo)	Kwa	14,700,000
Akan group	Kwa	8,600,000
Kongo	Benue-Congo	6,300,000
Zulu	Benue-Congo	5,500,000
Rwanda (Banyaruanda)	Benue-Congo	5,200,000
Xhosa (Xosa)	Benue-Congo	5,300,000
Luba	Benue-Congo	5,100,000
Shona	Benue-Congo	5,400,000

*Figures rounded off to the nearest 100,000.

Niger-Kordofanian

West Atlantic subgroup. The centre of gravity of the 43 West Atlantic languages would seem to lie in the Senegal–Guinea area, in which are found the two most important languages of the subgroup—Wolof and Fulani (the latter of which is also known as Fulbe, Fula, and Peul). The area is also the locus of the great majority of the other languages in the subgroup. Fulani, the language of a highly mobile pastoralist people numbering as many as 11,500,000, is also found in a scattered distribution as far east as Chad. The immense extension of this language, together with its association with cattle and Islām, led earlier researchers to give it a special place apart from the other West Atlantic languages—in fact, outside what is now recognized as Niger-Congo. It was first isolated in 1883 by an Englishman, Robert Needham Cust, who based his work on that of Friedrich Müller, an Austrian linguist. The German Karl Richard Lepsius classified it as one of the "mixed Negro" languages not fully sub-Saharan or Bantu (1880). Subsequently, it was included as a "Hamitic" language in the German Africanist Carl Meinhof's influential classification (1912), which was adopted by many scholars writing in English.

A feature of Fulani that is shared systematically with some of the other West Atlantic languages is an "alternation" whereby both the beginnings and the endings of words go through parallel changes according to grammatical considerations. This feature is found in its greatest elaboration only in Fulani; it is represented in either vestigial or undeveloped form in most of the other West Atlantic languages. This fact made Fulani appear more different from the others than it actually is and contributed to its misclassification. It is of interest to speculate why this most "vigorous" of the West Atlantic languages should be the most difficult or complex in its grammatical system. ("Vigorous" is used here in the sense that it has several millions of speakers, and also in the sense that it is associated with a dynamic expansionism). By comparison, Latin indicated grammatical case, gender, and number only by the endings of words, and English has virtually eliminated case and gender and has largely simplified number to a single suffix.

Bambara and Mende languages

Mande subgroup. The Mande subgroup includes 26 languages, of which the most prominent are Bambara of the Mali–Guinea area and Mende of Sierra Leone. In contrast to the classification of the West Atlantic subgroup, there has been little doubt as to the identification of the languages in the Mande subgroup. This certainty results from the subgrouping conforming closely to the "classic" expectation of language differentiation, in which a small number of contemporary languages are clearly seen to share a number of striking features not found elsewhere. Such features presumably provide evidence of early segregation of the ancestor language of the subgroup and of a continued effective isolation. Examples of such key features in the Mande languages are the systematic distinction between free and dependent nouns (see below) and the form and use of numerals.

As would be expected, however, the results of a longer period of independent development have also deflected the attention of many researchers to the differences of Mande from the Niger-Congo group—*i.e.*, to the common negative evidence of relationship with the rest of the Niger-Congo family. This negative evidence led Maurice Delafosse, in 1924, to propose two criteria for the identification of this subgroup: (1) a complete absence of the noun-class system that characterizes other West African languages and (2) a complete absence of tonal contrasts. Investigators who view Mande languages in terms appropriate to the group itself have shown that this is a prejudiced view and that a more positive formulation of these criteria is possible. Such a restatement would run somewhat as follows: (1) Instead of the particular distinctions among nouns reflected in the grammatical classes of the other subgroups of Niger-Congo, the Mande subgroup has elaborated a grammatical distinction between free and dependent nouns—nouns referring to "alienable" or "transferable" possessions, as opposed to "inalienable" possessions such as kinsmen or parts of the body. (2) The role of tone is tied to features other than those of semantic reference (*i.e.*, the meaning of the word taken in isolation).

Role of tone in the Mande languages

Whereas Chinese and other tone languages use changes of pitch to distinguish meaning in words with the same configuration of consonants and vowels (when pronounced by themselves), the Mande languages reveal a different use of tone, one that also differs from practically all the other Niger-Congo languages. In the Mande languages, a difference in meaning based on tone seems to demand a contrast of complete utterances, which are usually made up of more than one word. This contrast of entire utterances places Mande at one end of a continuum of distinctions based on tone, because linguists are still finding many variations in the situation of "meaningful semantic tone" or "contrastive tone patterns in isolation" that seem to demand some sort of context in order to be operative. In this respect, it is of interest that most of the differences in meaning between complete utterances that are identical except for tone in the Mande languages are more of a grammatical nature than of a semantic one. It must be stressed, however, that the Mande languages *are* tone languages and that the contrasts so produced are discrete and absolute and thus not of the "sliding" nature of speech melody. In form, they are closer to the question intonation of English, but in function they perform grammatical work internal to the sentence rather than coextensive with it.

It is striking that several independent writing systems based on the syllable (which is the unit that bears the tone) have been reported for Mande languages. the best known of these is the one devised for the Vai language, although Mende, Toma (Loma), and Kpelle also have indigenous syllabic writing systems. This type of writing is somewhat reminiscent of the writing system of Chinese (in which the syllable is often both the tone-bearing unit and the meaning-bearing unit) and of Cherokee, a North American Indian language.

Voltaic (Gur) subgroup. The Voltaic (Gur) subgroup includes 79 languages spoken in Burkina Faso (Upper Volta) and the upper halves of Ghana and Côte d'Ivoire (Ivory Coast). Mossi is the most prominent of the Voltaic subgroup. The term *Gur* was used as early as 1886 to designate the languages centred around Ouagadougou in what is now Burkina Faso and northern Ghana. Languages with such names as Gurma or Gurunsi were later related, and the first syllable of the names was taken as a name for the group. In 1911 Delafosse used the designation "Voltaic" for the same group, probably because many of the Gur languages are spoken in the Volta River basin.

Typological significance of the Voltaic languages

The study of the Voltaic languages as a group is important because they do not partake strongly of any entrenched typological or grammatical "manifestations" (such as the "manifestations" of sex gender distinctions in Indo-European, triconsonantal roots in Semitic, and so on). A wide range of possibilities is present, and, if any "manifestation" is at all apparent, it is a tendency toward simplification and serial constructions. This means that there are a very small number of closed grammatical classes and a greater reliance on constructional markers. Nevertheless, it is important, when considering any of the 1,000 or more languages surveyed in this article, to emphasize that each language within the subgroup is also important for its own sake. Each language reveals a particular way of communicating about the world, a way that is fully developed and elaborated in its own terms. These spoken languages are certainly not "underdeveloped" or primitive. In fact, the very complexity of analyzing the Voltaic languages (or any other group of natural languages) in terms and categories sufficiently "universal" to be operational both for those who already speak the languages and for those who want to learn something about them reflects their developed nature.

The most widely known characteristic of the Voltaic languages is the realization of the Niger-Congo noun classes in the shape of parallel prefixes and suffixes, in which the suffixes appear to be dominant. In languages like Mõõre and Dagbane, for instance, nouns for a "person" end with the suffix *-a* (*-u* in Kasem), which is replaced by *-ba* in the plural. This feature is somewhat reminiscent of Fulani of the West Atlantic subgroup but differs in that it typically appears in the form of separable syllables in the Voltaic languages, rather than as an alternation or mutation of the beginning and ending consonants of the root itself, as in Fulani. The partial classification of human beings, animals, and liquids into separate grammatical classes provides evidence for the original nature of such classes in the ancestral Niger-Congo language. So does the fact that this classification governs the participation of variant forms of the detachable suffixes (or suffix–prefix frames) in sentence constructions calling for pronounlike behaviour—referential, possessive, demonstrative, relative, and so on. Thus, Bargu has the suffix *-a* for the "person" class, a suffix that always "governs" the occurrence of *yé* as independent pronoun, *yù* as object marker, and *yè* as one demonstrative.

Ga, Yoruba, Igbo: prominent Kwa tongues

Kwa subgroup. The Kwa subgroup includes 73 languages, many of which are prominent (*e.g.*, Ga, Yoruba, Igbo). This grouping includes some of the most well known of the Niger-Congo languages and some with the greatest number of speakers. The Kwa languages are spoken in the belt of tropical forest through the southern section of the West African bulge. They are the languages of the former great kingdoms of tropical West Africa, such as Ashanti (Akan languages), Dahomey (Ewe), Oyo (Yoruba), and Benin (Edo, or Bini, language). In addition, the large decentralized group of Igbo also speak a Kwa language. Each of these languages has millions of speakers and a growing written literature. Yoruba is the language with the earliest published dictionary and grammar written by a native speaker. It is symptomatic of the position of these languages in modern life, however, that these Yoruba works were written in English by an African Anglican churchman, Bishop Samuel Crowther. English or French remains the national language of the countries in which Kwa languages are spoken, and interest in employing the local languages is largely restricted to religious or academic institutions based on European models.

Like the Voltaic subgroup, the Kwa languages are diverse in their structure and characteristic features, and examples of this diversity are found in the roles of tone and vowel harmony. The continuum from clear tones in isolated dictionary citations to modification in sentence context referred to in connection with the Mande languages is particularly striking in the Kwa languages. Thus, some investigators heard consistent pitch levels that could be noted on a musical staff for any one speaker (in at least one Kwa language, Grebo of Liberia, there are four such levels). Others, such as those investigating Akan, heard just as clearly that there is a complicated system of pitches that are varied according to transitions between syllables or interrelations among syllables. It is as yet unclear just which, if either, of these systems is historically earlier: whether the clear tone levels are a crystallization of the distinctions imposed by contextual reactions with other tones or whether the languages that have almost lost all independent lexical tone are a newer reinterpretation of a previous system of fixed pitches.

Vowel harmony in the Kwa tongues

Vowel harmony involves a system of mutual selection or restriction among the vowels that can go into the makeup of a word. The vowels in a word are said to be in "harmony," with high vowels usually occurring with high vowels, front vowels with other front vowels, and so forth, in such a way that a constant feature seems to be realized throughout the word. Vowel harmony of a type peculiar to sub-Saharan languages was termed "cross-height" vowel harmony by the British scholar John M. Stewart. In this type of vowel harmony, the higher of two varieties of a vowel selects the matching higher variety of another vowel and vice versa. Thus, *u* (as in "rule") would occur with *o* (as in "bone"), and the lower sound ʊ (as in "pull") with the correspondingly lower ɔ (as in "law"). Stewart also summarized evidence for contrast in the formation of the sounds in terms of the advancement of the root of the tongue. Under conditions of change of a vowel system, this leads to reinterpretations that cloud the nature of earlier contrasts and add to the variety found in the Kwa languages.

It is difficult to provide a generalized "profile" of the Kwa languages that would give an idea of what is typical of a language of this subgroup. In general the words are mostly short and the nouns begin with vowels. Proper names often appear to be quite long, however, because they are compounds or phrases made up of several words. Many words of Kwa origin are found in the Americas, as is the pantheon of Yoruba gods in South America and the Caribbean.

Adamawa-Eastern subgroup. The Adamawa-Eastern subgroup comprises 112 languages spoken in the Central African Republic and northern parts of Cameroon and Zaire. These geographically remote languages are the least known of the Niger-Congo family. The subgroup's most prominent members are Sango and Baya (Gbaya). The name Adamawa-Eastern implies a combination of two recognized sub-subgroups: Adamawa and Eastern. None of the language names of this subgroup is well known to outsiders, with the possible exceptions of the ethnographically noted Zande (Azande) and the trade language Sango.

Sango is clearly simplified from a dialect of one of the Adamawa-Eastern languages. It shows the result of contact with French as well as the usage of many tribes in trade along the Ubangi River. It is the lingua franca of the Central African Republic and is extremely widespread as a second language in Central Africa. Many other trade languages based on an African vernacular also are spoken in this central area of great rivers—*e.g.*, a pidgin dialect of Swahili, a pidgin form of Kongo (Kikongo), and Lingala.

Benue-Congo subgroup. The Benue-Congo subgroup comprises 557 languages, spoken from Nigeria to South Africa. Its most prominent members include Swahili, Kongo (Kikongo), and Zulu. In terms of numbers of speak-

Bantu, largest Benue-Congo subgroup

ers and geographical extension, Benue-Congo and Bantu are practically the same entity. That is, compared with the many millions of speakers of Bantu, the speakers of non-Bantu languages are only a tiny minority of Benue-Congo speakers. Furthermore, the Bantu speakers are spread out over most of middle and southern Africa, whereas the remainder of Benue-Congo speakers occupy only a corner of the northwest extension, mostly in Nigeria. The only two major non-Bantu Benue-Congo languages are Efik and Tiv, found in eastern Nigeria.

The following description sacrifices a consideration of the numerous but less important non-Bantu languages of the subgroup for a still superficial view of the Bantu languages. The most notable feature of the Bantu languages is their system of noun classes (concord). This has also been called "alliterative" concord because it often entails a repetition of the same syllable as prefixes at the head of successive words in the sentence. An example, drawn from Swahili, provides an illustration: *wa-tu wa-le wa-mefika* (noun-demonstrative-verb) "those people have arrived." Such prefixes as *wa* represent an intersection of two systems: one syntactic (function in the sentence) and the other paradigmatic (mutually exclusive replacements in the same position in the sentence). Syntactically, the shape of the prefix can vary according to the part of speech to which it is attached—a prefixed syllable may be different in the noun, adjective, possessive, locative, or verb. That there is agreement (*i.e.*, a concord element is present) with virtually all parts of speech creates an aesthetically pleasing system by means of an exhaustive display of marked relationships in the sentence.

Paradigmatically, the system of concord is most simply viewed as being based upon the governing noun in its singular or plural form. That is, the noun belongs to one of a number of noun classes that determine the proper set of concording elements to be used throughout the part of the sentence dominated by that noun. These sets (which can be thought of as replacement sets) come in pairs according to whether the governing noun is singular or plural; *e.g.*, in Swahili, the singular form *m-tu yu-le a-mefika* "that person has arrived" contrasts throughout with the plural form *wa-tu wa-le wa-mefika* "those people have arrived." This alternation of singularity and plurality is, of course, more closely related to the semantics of the noun than to its purely grammatical function. The paradigmatic system goes even further in this direction by making a rather loose use of other possible semantic contrasts (especially: person–nonperson and countable–noncountable) in assigning nouns to noun classes. In no language, however, is there a strict relation between meaning and noun class. Nevertheless, the tendency in this direction again stimulates a sense of completeness of aesthetic form and has added to the fame of Bantu languages as abstract figures.

Noun concord in Swahili

Correlates of these noun classes (in both form and function) are found in all branches of Niger-Congo except Mande and only slightly in Kwa. This led some early researchers to combine all of them as one subgroup and later led to a typological distinction between noun-class and non-noun-class languages. This distinction is not very useful, however, because there are a number of languages with rudimentary or partial noun-class systems, and it even seems possible for a language of entirely different ancestors to adopt noun prefix alternation.

Swahili remains the best known Bantu language. It is the national language of Tanzania and is used as a lingua franca throughout East Africa. The use of Swahili for this purpose is facilitated because it no longer makes use of tone—it is an isolated non-tone language in the Bantu area. There is another phenomenon related to tone in the Bantu languages besides its disappearance in Swahili that should be mentioned: there exists an area of "tonal reversal," a geographic area centred in the western Congo in which the tonal realizations in words and sentences are the mirror images of what they are elsewhere in Bantu languages; *i.e.*, high tones become low tones, and vice versa.

Area of tonal reversal

Another distributional characteristic is the presence of a prefix form made up of two syllables as opposed to the form with only one. Following out the implications of the geographic distribution as well as the function of this feature is of help in mapping the prehistory of the Bantu languages. As in all other areas of the world, however, it is the distribution of certain items of vocabulary, or of vocabulary replacements, combined with the application of laws of sound change, that produce the most valuable historical conjectures. An impressive amount of work has been done on the reconstruction of ancestral Bantu word roots and the sound system used to realize them. As early as the 1890s, Carl Meinhof postulated an original sound system and started working out the shifts that it had been subjected to in order to produce the various Bantu languages as they are spoken today. Meinhof called his systematic reconstruction Ur-Bantu, and most subsequent historical research (and much actual description of contemporary languages) has been based on his surprisingly trustworthy descriptions. *Comparative Bantu* (4 vol., 1967–70), by Malcolm Guthrie of the University of London, is a much fuller and more painstakingly detailed compendium of the entire geographical distribution of the Bantu languages.

Although textbooks and lessons have been published for many Niger-Congo languages, those of the Bantu area are by far the most numerous. In part, this availability reflects the influence of the South African linguist Clement Doke and the department of Bantu studies at the University of the Witwatersrand, which was in direct touch with a huge area of Bantu-speaking peoples.

ADOPTION OF FOREIGN WORDS AND SOUNDS

All the subgroups of Niger-Congo contain languages that show evidence of intensive contact with each other or with languages outside the group to such an extent that special mechanisms are in operation for the inclusion of borrowed words (such as the assignment of nouns to a particular noun class). This does not make them especially unusual from the standpoint of the world's languages, but the borrowing of the very difficult click consonants from the neighbouring Bushman-Hottentot languages into the Bantu languages of South Africa is definitely remarkable. Clicks are especially widespread in Zulu and Xhosa—the first consonant of even the name Xhosa itself stands for a click. This dramatic exception to the theory of language change that states that languages alter in order to simplify or economize the effort of sound production seems to require a special explanation. Possibly in-marrying Bushman or Hottentot wives used their native click sounds as disguised forms of words that would be taboo in Zulu or Xhosa. (D.W.C.)

Borrowing of click sounds

Chari-Nile and Nilo-Saharan languages

The term Chari-Nile refers to a group of languages presumed to be genetically interrelated, or descended from a common ancestral language. Joseph Greenberg proposed the Chari-Nile family as part of his overall classification of African languages. At first, he designated the group by the name Macro-Sudanic (1955), but in 1963 adopted the term Chari-Nile. The name Chari-Nile is geographically descriptive in that virtually all of the languages in the family are located in the watersheds of the Nile and the Chari rivers or in the areas in between them.

CLASSIFICATION

The larger Nilo-Saharan group. Greenberg, as of 1963, regarded Chari-Nile as only a branch of a larger family, Nilo-Saharan, of which it is by far the most heterogeneous. The other branches of Greenberg's Nilo-Saharan family are: (1) Songhai, an important language with no close relatives, which consists of a number of dialectal varieties (*e.g.*, Zerma) spoken along the Niger River in Mali and Niger; (2) Saharan, a language group including Kanuri, the major language of northeastern Nigeria, as well as its relative Teda and its more distant relative Zaghawa, both of which are spoken to the east of Kanuri, principally in Chad, but also in The Sudan; (3) Maba, a group of interrelated languages of Chad; (4) Koma (also written Coman), a group of interrelated languages of the Ethiopia–Sudan border area; and (5) Fur, a language of Northern and Southern Darfur provinces, The Sudan.

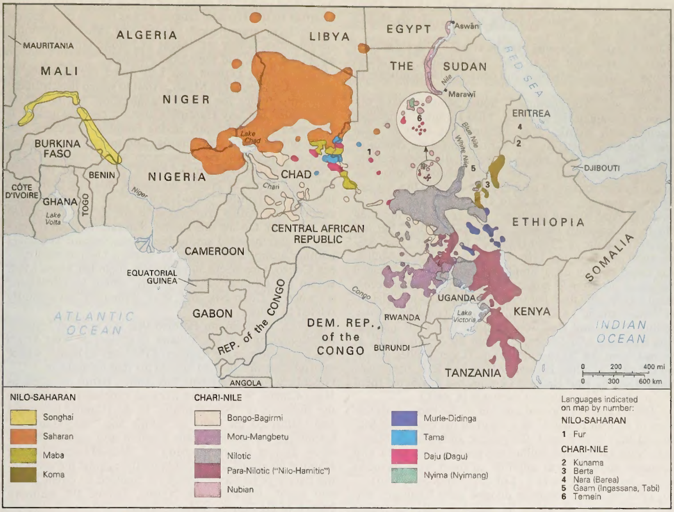


Figure 32: Distribution of the Nilo-Saharan and Chari-Nile languages.

Adapted from A.N. Tucker and M.A. Bryan, *Linguistic Analyses: The Non-Bantu Languages of North-Eastern Africa* (1966); Oxford University Press (for International African Institute), London

Languages of the Chari-Nile group. Greenberg considered Chari-Nile (a now controversial division) to be composed of four branches. Some later linguists favour four independent branches, eliminating the Chari-Nile node altogether. The first, Kunama, is a language of northern Ethiopia and western Eritrea, contiguous to Nara (Barea), an Eastern Sudanic language (see below); Berta, the second branch, is spoken in the Ethiopia–Sudan border area. Greenberg called his third division Central Sudanic. It is composed of a large and heterogeneous group of languages that extends over northwestern Uganda, the southern Sudan, northeastern Zaire, Chad, and the Central African Republic, although languages of other groups are interspersed. The British Africanists Archibald N. Tucker and Margaret A. Bryan (1956) regarded Central Sudanic as two distinct unities, which they hesitated to group together; one group is Bongo-Bagirmi, of which Bagirmi, Sara, and a number of other languages are found in Chad and the Central African Republic, while Bongo and its closest relatives are concentrated in the extreme south of The Sudan near the Zaire and Uganda boundaries. The second division of this group, according to Tucker and Bryan and others, is Moru-Mangbetu (like Bongo and Bagirmi, these are the names of individual languages), which is concentrated in northeastern Zaire but overlaps into Uganda. In addition, it includes Efe, Lendu, and Madi, which is linked to Moru. The fourth and most heterogeneous branch of Chari-Nile is Eastern Sudanic.

The subgroups of Eastern Sudanic

Eastern Sudanic consists of 10 subgroups:

(1) Nilotic extends from the southern Sudan through Uganda, Kenya, and northern Tanzania, although it is interspersed with languages of other groups, and includes the so-called "Nilo-Hamitic" languages. (This name is controversial; see below *Nilotic and "Nilo-Hamitic."*) The Nilotic subgroup includes Shilluk, Dinka, and Nuer in the southern Sudan, Acholi in northwestern Uganda, and Luo, in western Kenya on the eastern side of Lake Victoria. The Nilo-Hamitic languages include Bari and Lotuko (Lotuho) in the extreme south of The Sudan; Karamojong in northern Uganda (to the east of Acholi) and its neighbour and close relative, Turkana, across the border in Kenya; Nandi and Suk, south of Karamojong; and Masai (Maasai; also called Maa) extending from southern Kenya into northern Tanzania.

(2) Nubian is spoken along the Nile from Aswān, Egypt, to Marawī, The Sudan. A group of Nubian languages sometimes known as Hill Nubian are spoken chiefly in the Nuba Mountains in the southern Kordofan region of The Sudan. Nile Nubian is divided by some scholars into three languages, though the northernmost, Kenuzi (Kenzi), and the southernmost, Dongola, are closely related, almost to the point of mutual intelligibility; Mahas, located between the two, is a more distant relative and includes a subdialect, Fadidja, spoken to the north.

(3) Murle-Didinga designates a group of closely interrelated languages spoken on both sides of the Sudan–Ethiopia boundary; Murle and Didinga are individual languages within the group.

(4) Nara (Barea), a language of western Eritrea, is spoken just to the north of Kunama.

(5) The Tama group of languages, spoken on both sides of the Chad–Sudan border, are closely interrelated. They include Merarit (Mararit) and Tama, for which the group is named.

(6) Gaam (also called Ingassana or Tabi) is found in the eastern Sudan near Lake Ar-Ruṣayriṣ.

(7) Daju, or Dagu, is the name of a group of closely related languages spoken in widely separated enclaves in Chad and The Sudan.

Later additions to the Eastern Sudanic branch

Three subgroups of Eastern Sudanic were added by Greenberg in 1963; the first two had previously been treated as entirely independent language families, while the third had not been mentioned. These are: (8) Nyangiya, also known as Teuso, a group of dialects or closely related languages of northern Uganda, whose inclusion in Eastern Sudanic Greenberg regarded as only tentative (this is omitted on the map); (9) Temein, a language of the Kordofan region of The Sudan; and (10) Nyima, or Nyimang, also found in the

Kurdufān region. Another language, Afitti, spoken about 100 miles (160 kilometres) to the north, evidently should be grouped with Nyima.

Controversies concerning classification. There has been much disagreement about the classification of various groups of African languages and considerable confusion in terminology.

Nilotic and "Nilo-Hamitic." Some scholars have objected to grouping the so-called "Nilo-Hamitic" languages together with Nilotic. This question has been at issue since the earliest attempts at African linguistic classification. One of the first proposed systematic classifications (1877–78) established the very heterogeneous Nuba-Fula group, named for two of its members, Nubian (an Eastern Sudanic language, according to Greenberg) and Fula, or Fulani (a West African language). The group also included Masai, subsequently classified as "Nilo-Hamitic," but not the other Nilotic or "Nilo-Hamitic" languages. This classification, made in the 19th century by Friedrich Müller, was based mainly on a simplistic correlation with racial categories determined largely by hair type. Müller recognized a close affiliation between Bari (subsequently classed as "Nilo-Hamitic") and Dinka, a Nilotic language. Dinka, Shilluk, Nuer, and Bari were grouped under the heading Nile languages (*Nil-Sprachen*) within the Negro group, as distinct from the Nuba-Fula group. (Later Müller admitted the close affinities between Masai and Bari, previously pointed out by Karl Richard Lepsius, but attributed them to a mixture of two distinct elements in Bari.)

Objections to grouping Nilotic with "Nilo-Hamitic"

In 1880, Lepsius proposed another of the early classifications of African tongues. As a result of his views that Masai and Bari were related, the genetic affiliation of Masai, Bari, Nuer, Dinka, and Shilluk was generally accepted, and languages showing clear affinities to them were included as they became known (*e.g.,* Lotuko, Nandi-Suk, Luo).

In 1902 Sir Harry Johnston divided these languages into two groups, one of which he called Nilotic, including Dinka, Shilluk, Acholi, Alur, Lango, and Luo. He proposed no specific name for the second group, which later came to be known as "Nilo-Hamitic," but included in it Masai, Turkana, Lataka (*i.e.,* Lotuko), Nandi, Suk, and Karamojong. He posited a distant relationship between the two groups but also noted some striking lexical similarities between Somali (a Hamitic language) and the second, or "Nilo-Hamitic," group that led him to conclude that the latter resulted from a mixture of one group of people related racially and linguistically to the Somali and another related to the Nilotes.

Most investigators of this period grouped these languages in much the same way as Johnston; for example, the English diplomat and scholar Sir Charles Eliot stated:

Correspondence of Sir Charles Eliot's classification with Nilotic and "Nilo-Hamitic"

> East Equatorial Africa is the home of a group of tribes which cannot be called Bantu or Hamitic, and have received no satisfactory designation. The group contains at least two subgroups. One of these, including the Shilluk, Dinka, Bari, Acholi, and Jaluo [*i.e.,* Luo], inhabits the banks of the Nile and the shores of Lake Victoria; the other has its headquarters in the highlands of . . . East Africa. . . . Its chief members are the Masai, Nandi, Turkana, and Suk. . . . To these should probably be added Latuka [*i.e.,* Lotoho] but our information about it is slight. . . . Karamojo is not well known but is apparently closely allied to Turkana.

Except for Bari, which was put in the first group for reasons more geographic than linguistic, but which is actually closely related to Masai, Eliot's two divisions correspond exactly to Johnson's—namely, Nilotic and what later became known as "Nilo-Hamitic."

A sharply different view was introduced in 1912 by Carl Meinhof, who denied the linguistic unity of the two groups suggested by Johnston and Eliot. Meinhof regarded the term Nile (or Nilotic) merely as a geographic concept. He clearly considered Bari and Masai to be Hamitic languages (*i.e.,* related to Somali and Galla, among others) and also, presumably, Lotuko, Turkana, Nandi, and Suk. (He was not, however, explicit about these.) Dinka and its close relatives—Shilluk, Nuer, Luo, and others—were classified as Sudanic; the resemblances between the two groups (in particular those between Dinka and Bari, which had been pointed out by Müller) he attributed to borrowing. Meinhof's most important reason for allying Masai and its close relatives to the Hamitic languages was that in both groups nouns have grammatical gender and other inflectional complexities, which are absent in Nilotic languages such as Shilluk, Luo, and their close relatives.

Much earlier, Lepsius also had noted the presence of grammatical gender in Masai, a trait to which he attributed great importance for genetic classification, but he specifically ruled out any Hamitic affiliation because gender is expressed very differently in the Hamitic languages. Because of Meinhof's great prestige and preeminence among Africanist linguists, his views were widely accepted. The scholar Dietrich Westermann (in 1912), however, regarded all of the above as Nilotic languages, dividing them into the Niloto-Hamitic (Nilo-Hamitic) and Niloto-Sudanic subdivisions, which correspond to the groups delineated by Johnston. On the other hand, contrary to general practice, he also included among the Niloto-Sudanic languages another group, Moru-Madi, of western Uganda and eastern Congo. Greenberg subsequently included Moru-Madi within the Central Sudanic family, which he considered a branch of Chari-Nile and, thus, ultimately but distantly related to Nilotic.

In 1935, Westermann claimed that the relation between Niloto-Sudanic and the Sudanic languages of West Africa was not as close as the relation of the latter to Bantu, whose interrelationship he had sought to demonstrate in 1927. Greenberg was to designate this family, Bantu included, as Niger-Congo. Westermann likewise expressed doubt as to the affiliation of "Nilo-Hamitic" to the Hamitic languages and reaffirmed the unity of Nilotic and "Nilo-Hamitic." Greenberg also adopted this position and argued persuasively against the relationship of "Nilo-Hamitic" to the Hamitic languages and, consequently, against the misleading nature of its name, which he changed to Great Lakes and which, together with Nilotic, formed the southern branch of his Eastern Sudanic family.

Doubtful affiliation between "Nilo-Hamitic" and Hamitic

The German Africanist Oswin Köhler proposed, in 1955, the name Nilotic for the entire southern branch and divided it into three subdivisions: West Nilotic (including Nuer, Dinka, Shilluk, Luo), East Nilotic (including Bari, Masai, Lotuko, Karamojong), and South Nilotic (including Nandi, Suk, Tatoga). (The last two had usually been grouped together as Nilo-Hamitic, while the first had simply been designated Nilotic.) In 1963 Greenberg subsequently accepted Köhler's grouping and nomenclature. Tucker and Bryan, while still maintaining the unity of "Nilo-Hamitic" and hesitating to group it unequivocally with Nilotic (in the older restricted sense, equivalent to Köhler's West Nilotic), proposed a new name for the group, Para-Nilotic, thus avoiding any implication of Hamitic affiliation.

The wider relationships of the Nilotic family. The question of the relationship of Nilotic (in Köhler's extended use of the term) to other African languages and language families has been raised at various times. In 1920 the English anthropologist G.W. Murray noted similarities between Bari and Nubian and, in fact, between Nubian and the Nilotic and "Nilo-Hamitic" languages as a whole. He also linked these languages to Kunama, Barea, and the recently discovered Tabi (also known as Ingassana). Westermann suggested in 1935 that Didinga might be a distant branch of Nilotic and had earlier (1912) made a similar claim for the Moru-Madi group. All these languages were subsequently classed by Greenberg either as Eastern Sudanic or within the wider grouping of Chari-Nile.

Apart from Eastern Sudanic, the most heterogeneous subgroup of Chari-Nile is Central Sudanic. Its unity was first suggested in 1940 by Tucker, who included it in a group of languages that he called Eastern Sudanic, a term that he admitted was essentially geographic because it included another group of different and distinct languages, among them Ndogo-Sere and Azande. (Greenberg subsequently incorporated these into Niger-Congo.) Greenberg introduced a terminological confusion here by calling part of Tucker's Eastern Sudanic group Central Sudanic and using the name Eastern Sudanic for an entirely different group of languages. Although Tucker hesitated to unite the Moru-Madi and Bongo-Bagirmi languages—which together comprise Greenberg's Central Sudanic—into a single family, the evidence that he himself presented appeared to support such a grouping.

Opposition to Greenberg's groupings

Some scholars doubt that adequate evidence has been presented for regarding the whole of Eastern Sudanic as a valid subdivision of Chari-Nile or even for regarding the whole of Chari-Nile as a valid subdivision of Nilo-Saharan. Although Nilo-Saharan and the more heterogeneous subdivisions within it are far from conclusively established, no convincing alternatives have been suggested.

LINGUISTIC CHARACTERISTICS

The Chari-Nile languages are a very heterogeneous group whose structural features, both phonological and grammatical, are in no way uniform or easily characterized. Furthermore, the great majority of the languages are not adequately described, and their comparative study is still in its infancy.

Phonology. Nearly all Chari-Nile languages seem to be tonal (*i.e.*, words and grammatical forms are differentiated by pitch), as are the overwhelming majority of sub-Saharan languages. Some varieties of Nubian may be exceptions, but the question has yet to be fully investigated. Central Sudanic languages generally have only syllables that are open (*i.e.*, end in vowels); some final vowels in the Bongo-Bagirmi languages of this division, however, are semi-mute or whispered. Other Chari-Nile languages have syllables that are open as well as closed (*i.e.*, end in consonants). Central Sudanic languages likewise differ from other Chari-Nile languages in that they are the only ones with the labiovelar consonants *kp* and *gb* (articulated with both the lips and the soft palate), which are common in much of West and Central, but not East, Africa. As in other African language families, a large number of Chari-Nile languages distinguish dental from alveolar or retroflex stops, among them most Central Sudanic languages, Nyima, Temein, Murle-Didinga, Hill Nubian, and West Nilotic. (The dental stops, *t* and *d*, are pronounced with the tongue tip against the back of the upper teeth; the alveolar and retroflex stops are pronounced with the tongue tip farther back along the roof of the mouth.) Some Chari-Nile languages—most Central Sudanic languages, Daju, the Murle-Didinga group, and Bari (East Nilotic)—distinguish implosive and explosive voiced stops; the former are pronounced by drawing the air into the mouth, the latter by expelling it. The vowel systems of most Chari-Nile languages are fairly rich, the great majority seeming to distinguish seven or more vowel qualities. Apparent exceptions are Daju, Nile Nubian, and possibly Kunama, which have only five vowels. Many languages make distinctions in the length of vowels (*e.g.*, Nile Nubian, Nilotic, Daju, Kunama), but not all (*e.g.*, Central Sudanic, Nyima). In Nilotic languages a contrast exists between so-called breathy versus hard vowels, sometimes analyzed as close versus open. Nasal vowels occur in Sara-Mbai of the Bongo-Bagirmi group within Central Sudanic but in no other Chari-Nile language.

Vowel systems of the Chari-Nile languages

Grammar. No Chari-Nile language has concordial noun classes like those of Bantu (see above *Benue-Congo subgroup*) and many of its more distant relatives in West Africa, but grammatical gender is distinguished in three very different Chari-Nile groups. It is prominent in East Nilotic (*i.e.*, "Nilo-Hamitic" minus the Nandi-Suk-Tatoga group) but absent in the other Nilotic languages. There are two genders of nouns, normally indicated by a determiner (in Masai, these are *l* for masculine, *n* for feminine) with various grammatical functions. On the other hand, gender is not distinguished in the personal pronouns, but in Daju (grouped by Greenberg in Eastern Sudanic) there are three genders (masculine, feminine, neuter) distinguished in the 3rd person singular pronoun ("he," "she," "it") but not in the noun. The same usage occurs in Bongo but in no other Central Sudanic language.

A few Chari-Nile languages distinguish exclusive forms from inclusive forms of the 1st person plural pronoun. The exclusive pronoun excludes the hearer (*e.g.*, "he and I" = "we"); the inclusive includes him (*e.g.*, "You and I" = "we"). Among such languages are Lendu (alone among Central Sudanic languages), Daju, most West Nilotic languages, and Teso (alone among East Nilotic languages). Virtually all Chari-Nile languages except Berta and Nyima formally distinguish plural from singular nouns, some by means of prefixes (*e.g.*, the Moru-Mangbetu languages within Central Sudanic, and Temein) but most with suffixes (*e.g.*, the Bongo-Bagirmi languages within Central Sudanic, Kunama, Barya, Daju, Ingassana, Murle-Didinga, Nyangiya, and East and South Nilotic, which together comprise "Nilo-Hamitic"). West Nilotic utilizes suffixes sometimes, as well as change of tone, vowel length, or vowel quality. These processes can also be found to some extent in other branches of Nilotic. Affixes are sounds or groups of sounds that are added to the beginning (prefixes) or end (suffixes) of a word or inserted in the middle (infixes). In Chari-Nile, it has been noted that two pairs of singular-plural affixes are widespread, though not universal; these are also used to some extent outside this family. The affixes are *t* and *k*, confined principally to nouns, and *n* and *k*, principally used with pronominal elements. The *t* and *k* forms occur in fewer languages than the *n* and *k* ones, and only in languages in which the *n* and *k* are also found. The significance of these grammatical features for the historical and comparative study of the Chari-Nile languages remains to be investigated.

Widespread use of specific affixes in Chari-Nile

Writing. Only one Chari-Nile language, Nubian, has a written tradition of any antiquity. Dating from the 8th to the 11th centuries, the script was employed by Christianized Nubians in much the same area as that occupied by modern Nile Nubian speakers. Generally called Old Nubian, this written language most closely resembles Mahas among the modern dialects. It was abandoned after the Nubians adopted Islām and survives only in ancient manuscripts and inscriptions that were not deciphered until 1906. (Since the adoption of Islām, Nubian has occasionally been written in Arabic script.) Old Nubian utilized a script derived from that of Coptic, which in turn was adapted from the Greek alphabet. Coptic added seven non-Greek letters of Egyptian demotic origin to represent non-Greek sounds. Nubian retained three of these letters, in addition to adding three new ones from cursive forms of Meroitic letters.

Meroitic, the extinct language of an even more ancient Sudanese civilization (*c.* 300 BC to AD 100), may be related to the Chari-Nile languages, but too little is known to establish any conclusive tie. It is not closely allied to any existing language. Surviving only in inscriptions, Meroitic was written in a consonantal alphabet derived from Egyptian hieroglyphic writing and in a cursive form partly of Egyptian demotic and partly of indigenous origin. In the first decade of the 20th century the script was deciphered by the English Egyptologist Francis Griffith.

Languages using Latin-based alphabets

Since the colonial period a number of Chari-Nile languages, particularly in Kenya, Uganda, and The Sudan, have been written in derivatives of the Latin alphabet and are used for religious and educational purposes. Nearly all of the languages are Nilotic. Of these, Luo, a West Nilotic language and one of the major indigenous languages of Kenya, has the largest vernacular literature. It also has the largest number of speakers of any Nilotic tongue (and, in fact, of any Chari-Nile language). Other West Nilotic languages, however, such as Dinka and Nuer of The Sudan and Acholi of Uganda, have some vernacular literature, as does Kalenjin (Nandi-Kipsigi), a South Nilotic language of Kenya. (M.F.Go.)

Khoisan languages

The Khoisan languages are click languages spoken in southern Africa. The term Khoisan was created to refer to the related peoples known as Bushmen and Hottentots (*i.e.*, the Khoisanid peoples) under a common name and has become increasingly accepted since its creation in 1928. The word is derived from Khoikhoi and San, the names of the peoples called, respectively and pejoratively, Hottentots and Bushmen.

The languages of the group are now usually divided into three groups—North, Central, and South Khoisan. They are spoken by remnants of a pre-Bantu population, the so-called Bushmen, who are hunters and collectors liv-

Table 59: Notation of Clicks

click formation type	characters used in text	International Phonetic Alphabet	Zulu conversion
Dental	/	ʇ	c
Alveolar	≠	o	
Alveopalatal (or retroflex)	!	ʗ	q
Lateral	⫽	ʖ	x
Bilabial	⊙		

ing in and around the Kalahari. A Khoisan language also is used by the Khoikhoin, pastoralists of Namibia. The Bergdama, a non-Khoisanid people, and the Hai-⫽'om San also have adopted the Nama language, a circumstance that brings the number of Nama speakers to a total of 146,000. The Khoisan family comprises about a dozen languages and dialect clusters such as Nama, !Kora, Naro, /Kham, !Khung, Kxoe, and others. (The unusual symbols in the names of the languages stand for clicks; they are listed in Table 59 and are explained below in the section on phonology.)

Problems of Khoisan genetic relationship

The assumption of a genetic relationship between all the languages of the Khoisanids and the application of the term Khoisan to this language "family" are problematic. Criteria brought forth against the point of view of a genetic relationship are the cleavage in morphology and vocabulary between sex-gender languages (Central Khoisan) and nongender languages (North Khoisan and South Khoisan), on the one hand, and a similar cleavage in morphology and vocabulary between North Khoisan and South Khoisan, on the other. These arguments are countered by the fact that there are a number of common words and a few particles, traceable by comparison of sound and meaning, in two or more of the three groups of Khoisan. These permit the establishment of a strong hypothesis for a genetic, though remote, relationship. Regular sound correspondence, however, has not been found in all common items; this is the most difficult part of the Khoisan problem.

The Khoisan hypothesis is built on: (1) common special features of the phonological system—*e.g.,* clicks are regarded as inherent to the languages; (2) widespread and common patterns of root formation, combined with special patterns of consonant distribution; (3) the occurrence of some probably related particles in more than one group; and (4) the occurrence of related words (sound-meaning units) in two of the groups, but more rarely in all three groups—*e.g.,* the terms for "chin," "lungs," "throat," and "wound."

In addition to Khoisan proper, there are two click languages, Sandawe and Hadza (Hatsa), spoken by peoples in Tanzania; these languages possess a few words, affixes, and particles that justify the assumption of a distant relationship with Khoisan. The apparent link between Central Khoisan and Sandawe is supported by the racial affinities of the Sandawe and Khoikhoin peoples. The relations between Khoisan and Hadza are more remote. Studies on this subject are still in the initial stages. The American linguist William E. Welmers proposed the tentative extension of the Khoisan family into Macro-Khoisan, including Sandawe and Hadza. In 1962 Joseph Greenberg subsumed Khoisan proper and Sandawe and Hadza under the title Khoisan, but a few later linguists continue to challenge this grouping.

Some Central Khoisan gender affixes were at one time considered as possibly related to the Hamitic languages. Today, the basis for the Nama-Hamitic hypothesis is

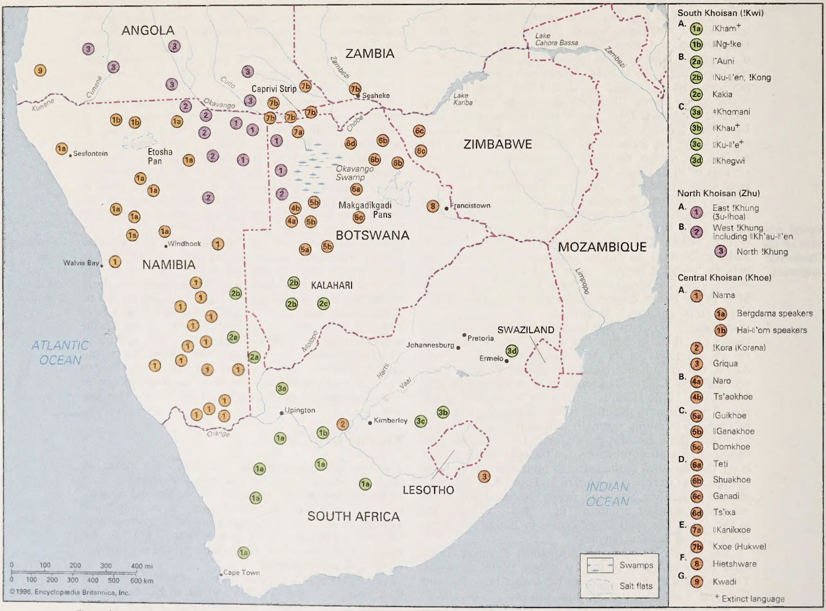


Figure 33: Tentative distribution of the Khoisan languages.

generally regarded as too narrow for conclusiveness and the hypothesis has been given up for lack of evidence.

CLASSIFICATION

All classifications of Khoisan are based on a general comparison of morphology and vocabulary. Lack of sufficient language material limits systematic comparative studies. While the division into groups is fairly consistent, the classification within the groups must be regarded as preliminary. The first classification of these languages, made by the German linguist Wilhelm Bleek in 1858, distinguished between the Hottentot and Bushman languages according to the gender-nongender dichotomy. This remained the traditional model until 1927, when Bushman was subdivided into Northern, Central, and Southern groups. These groups were maintained in D.F. Bleek's *A Bushman Dictionary* (1956), in which the Hadza language was included in the Central Group.

Modern theories of classification

Since 1950, several attempts at a reclassification of the languages of the Khoisanids have been proposed. The contribution of Greenberg to the Khoisan hypothesis was to include Hottentot in the Central group and, as mentioned above, to extend Khoisan to include the more closely related Sandawe and the more distantly related Hadza. Greenberg's main aim was to trace far-spread common elements in Khoisan as evidence for the genetic relationship of the groups. The method is based on sound-meaning comparison. The South African scholar E.O.J. Westphal separated the languages of the Bushmen from the gender languages that he interpreted as of Hottentot origin. Later, he divided the languages of the Khoisanids into five genetic groups, and in 1963 he further revised his classification of the non-Hottentot tongues by establishing four genetically unrelated groups, to which he gave the status of language families (1965); in 1971 he reduced these to three families. A new classification of Central Khoisan was put forth by Oswin Köhler in 1962, which showed the special position of the Nama and !Kora dialects of the Hottentots within the Central Khoisan group (see below).

There is no general agreement on the division of the South Khoisan languages. Some special phonological and morphological features characterize South Khoisan as a genetically related group of languages whose divergences must go back to a much older separation than that in Central and North Khoisan. South Khoisan speakers were probably the first Khoisanid migrants to reach South Africa.

The following is a classification, proposed by Köhler, of the languages subsumed in the three major Khoisan groups (a comma between languages indicates a close relationship; a semicolon designates a more distant relationship):

South Khoisan: /Kham, ⫽Ng-!ke; /'Auni, Xatia, /Nu-⫽'en, !Khong, /Namani, Kakia; ≠Khomani, /Nhuki, ⫽Khau, ⫽Ku-⫽'e, Seroa, ⫽Khegwi (Batwa), !Gãne.

North Khoisan (dialects): East !Khung (ʒu-/hoa), West !Khung (called ⫽Kh'au-⫽'en or Auen in its southern area), North !Khung (!'O-!Khung in Angola), closely related to West !Khung.

Central Khoisan: Nama, spoken by the Bergdama and by the Hai-⫽'om; !Kora (Korana); Griqua; Naro, Ts'aokhoe; /Guikhoe, ⫽Ganakhoe, ≠Heβa-khoe, Domkhoe; Teti, Danisa, Hura, Shuakhoe, Ts'ixa; ⫽Kanikxoe, Kxoe (Hukwe, Kwengo, Bumakxoe, including Bugakxoe); Hietshware, /Haitshuari, Mohisa; Kwadi (tentatively classified as Central Khoisan by Köhler in 1968 and Westphal in 1971).

Sandawe has similarities to Central Khoisan and may be regarded as distantly related. Hadza shows some relationships to Khoisan and relatively few connections to Sandawe. Its inclusion in Khoisan is problematic and assumes a very high, abstract level of classification, presupposing that the languages separated a long time ago and that a great number of cognates and common features that normally attest to a common origin were lost over an extended period. This high level of classification is also characteristic of all three Khoisan groups.

LINGUISTIC CHARACTERISTICS

Phonology. Most characteristic of the Khoisan languages are the click sounds, which form a subsystem of the consonants. Clicks are produced by simultaneous closure of the tongue with two areas in the mouth, one at the soft palate (the velum), and the other at such prevelar locations as the teeth, alveolar ridge, or palate. After the closures are made, the body of the tongue is moved down and back, so that the enclosed part of the mouth cavity becomes larger. This causes a lowering of the air pressure in the cavity. The click noise occurs when the tip of the tongue is lowered, breaking the partial vacuum and allowing air to rush into the cavity.

Click sounds as a subsystem of the consonants

There are two types of release of clicks. In one the bilabial, dental, and lateral closures are released by an imploding affricate (a sound beginning as a stop and ending with more or less distinctly audible friction); the alveolar and alveopalatal (or retroflex) closures are released by a hard implosive sound. Both releases are called ingressive, or influx, releases, meaning that the air rushes into the mouth. The second type is called egressive, or efflux, release; it results from the velar closure. In this case the air rushes out of the mouth. A click plus efflux are considered as a single sound rather than as a consonant cluster. In the South Khoisan languages there is also a bilabial click with labial closure; this is like the sound of a kiss. The Khoisan click mechanisms also combine with other phonetic features—*e.g.*, voicing (vibration of the vocal cords), nasalization—to produce additional sounds. The clicks occur only in initial position in word roots (also in compound roots), rarely in affixes and particles. Some of the Khoisan click sounds were adopted into Bantu languages, notably the Nguni group that includes Zulu and Xhosa (see Table 59 for click notations).

Unless followed by a vowel, the effluxive part of a click is represented by a separate symbol. In this text, the symbols are follows: /a = dental click plus vowel; /h = dental click plus aspiration; /x = dental click plus velarization; /x' = dental click plus velar ejected efflux; /' = dental click plus glottalization; and /n = dental click plus nasalization. According to this system, as well as in the general orthography of personal and place names, *k* or *g* is used to indicate the velar closure of the click to make the words readable for the general reader who omits the initial click sound. In this case *k* is used to indicate an unvoiced click and *g* a voiced click. The examples given above thus read as follows: /a = /ka or /ga; /h = /kh or /gh; /x = /kh or /gh; /x' = /kh' or /gh'; and /' = /k' or /g'. In the case of nasalization (/n), *k* and *g* are not usually written.

In the Khoisan languages, there are intricate systems of tones. The main vowels are *a, e, i, o,* and *u;* ɛ as in "bed" and ɔ as in "all" do occur, but not in all languages, and the *ə* sound, as the *a* in sofa, is relatively rare. Nasalized *ã, ĩ,* and *ũ* are most common. Pharyngealized (or pressed) vowels and vowels with intermittent glottalization occur in North Khoisan and, to a much lesser degree, in South Khoisan but seldom in Central Khoisan. (A pharyngealized vowel is one produced with constriction of the pharynx, the area at the back of the mouth before the esophagus.) A characteristic feature of the consonants is their use in clusters with final velarization (articulation of the back of the tongue with the velum, or soft palate), aspiration (an accompanying puff of breath), or glottalization (accompanying closure of the glottis). This occurs in all three Khoisan groups.

Grammar. Because the grammatical features of the various Khoisan languages are so divergent and lack overall uniformity, they will be discussed in three groups—South Khoisan, North Khoisan, and Central Khoisan.

South Khoisan. The languages of the South Khoisan group are characterized by a great number of different grammatical elements and processes. For example, the plurals are formed not only by the addition of suffixes but also by reduplication (*i.e.*, the repetition of an element to indicate plural, as "bird-bird" for "birds"), by partial change of the stems, and by completely different stems, called suppletive forms. In the formation of tenses and moods, a variety of particles occurs before the verb. Significantly, there is no class or gender distinction in the South Khoisan languages, except for a few traces in the /Kham language. The normal sentence order is subject-predicate-object.

Types of plural formation

North Khoisan. The North Khoisan dialects are characterized by an almost complete absence of affixes (endings, prefixes, and so on). The relations of the words to one another and their function in particular utterances are designated by particles (similar to English prepositions such as "to," "for," or "with") and by word order. The tense of roots functioning as verbs is shown by an optional preceding particle for present and future time, by nothing for the past tense, or by an expression describing time ("now," "tomorrow," etc.) if used in a sentence without context. As in the South Khoisan languages, there are suppletive (different) stems for some plurals; *e.g.,* *!hũ* means "to kill one," but *!'oa* means "to kill many." The North Khoisan dialects, also in common with the South Khoisan tongues, lack a passive construction and have the basic sentence order of subject–predicate–object.

Central Khoisan. In the treatment of the Central Khoisan languages, the Kxoe language of the west Caprivi Strip (Namibia) will serve as representative. All of the Central Khoisan languages have a highly developed morphology. Kxoe adds suffixes that indicate gender—masculine, feminine, and common in the dual and plural forms, and masculine, feminine, and neuter (*i.e.,* undetermined, with affix *a* or no affix) in the singular—to roots that take noun (nominal) endings. The same gender suffixes are added to the third-person pronouns (originally demonstratives). In grammatical gender, masculine implies "long, narrow, strong," and feminine suggests "short, round, broad, weak." For example, */u* or */u⊙ma* "canoe" is masculine, while */u⊙hɛ* "ferry" is feminine. Gender distinction is also found in the first persons (except singular) and second persons.

Masculine and feminine gender

There are six tenses for the verb, with the present tense also expressing whether the subject is standing, sitting, or lying. Unlike the other Khoisan groups, the Central Khoisan languages have a passive form, which is frequently used. In sentence order, subject–object–predicate is the most usual, but subject–predicate–object is also common, and object–subject–predicate occurs if the subject is a pronoun or a nominal.

Vocabulary. Khoisan vocabulary is highly adapted to the needs of life in a poor natural environment, especially to the hunt, to the gathering of food, and to all skills serving the preservation of life. As a result of contact with other cultures, loanwords have been adopted—*e.g.,* from European languages into Nama, from Bantu into Kxoe. Native Khoisan creations that reflect non-native innovations also occur—*e.g.,* Kxoe */am-mũ-'o-xo* "clock, watch" is literally "sun-see-on-thing." The use of words in a figurative sense to label new objects is another process for word-formation. Examples include Kxoe *yɛ-≠'am* "bridge," which is literally "hole upper side," and *kuru* "to drive [a car]," which derives from "to press the bellows." Gender symbolism also plays a role in the formation of words for objects in the contact cultures; *e.g.,* the Nama masculine term */nū-b* means "leg," while the feminine counterpart */nū-s* means "wheel."

Basic numbers exist for 1 and 2 and, in some languages, for 3; higher numbers, however, are circumscribed. For example, the number 4 in Kxoe is literally "lick hand bone quantity"—that is, "the finger with which one licks out a pot." The numbers in the languages of the Bushmen seldom go beyond 10; this contrasts greatly with the practice in Nama, and in Khoisan languages in general, in which there is a decimal system with noncompound numerals from 1 to 10.

In general, the unit of the vocabulary equivalent to the word in European languages is a root whose category as a noun or verb is determined by the context and by the use of noun and verb affixes or particles. Compound words are often used to extend meaning, as English-derived words do (*e.g.,* in Kxoe, *kx'â-xò,* literally "drink [verb]-thing" means "beverage").

Writing and texts. About 10 Khoisan languages have been recorded with various degrees of intensity by missionaries and linguists. The best known are Nama, Kxoe, !Khung, and /Kham, followed, to a lesser degree, by !Kora, ≠Khomani, ⫽Khegwi, Naro(n), and /'Auni. The old individual methods for recording clicks were superseded by the Lepsius system (1854), which is, with some modifications, still in use. Studies in Kxoe, !Khung, !Khong, /Guikhoe, and Sandawe have been carried out in recent times, with current research concentrating on !Khung. In view of the difficulties of research among vanishing bands of peoples, the prospects for gathering sufficient material in the remnant Khoisan languages are not good.

Folklore collections

The largest collection of folkloristic texts was taken down in the now extinct /Kham language of the former province of Cape of Good Hope. Next to /Kham, the most comprehensive collection of texts on history, folklore, and traditional and modern life was recorded in the Kxoe tongue. Texts on folklore and history also exist in Nama and Korana and to some extent in !Khung. Literacy is found only in the Nama language, among the Nama and Bergdama. In !Khung, the first two primers were printed in 1969. (O.R.A.K.)

LANGUAGES OF THE AMERICAS

Eskimo-Aleut languages

Eskimo and Aleut, once neighbour languages on the Alaska Peninsula, are related but quite distinct; together they form the Eskimo-Aleut language family. The proposed relationship of Eskimo-Aleut with other language families, such as Chukotko-Kamchatkan, Uralic, and/or Indo-European, remains conjectural.

CLASSIFICATION AND DISTRIBUTION

Clarification of names

Eskimo is a blanket term for Inuit and Yupik, the two mutually unintelligible main divisions of the Eskimo languages. The name Aleut, used by Russian fur traders from the Kamchatka Peninsula in 1745, refers primarily to the people of the Aleutian Islands, who call themselves Unangan or (in the dialect of Atka) Unangas, but also by extension to the Pacific Yupiks, who call themselves Alutiit (plural of Alutiiq), an adaptation of the Russian name.

Inuit. Inuit, which means "people" (plural of *inuk,* "person"), is used as a name for the language spoken in Greenland, Arctic Canada, and northern Alaska, U.S., west to the Bering Strait and south to Norton Sound. It is a dialect continuum, in which neighbouring dialects are mutually intelligible but the cumulative differences impede or prevent understanding between groups that are some distance apart.

Yupik. Yupik, a dialectal form meaning "real person," includes five languages: Central Alaskan Yupik, spoken southward from Norton Sound; Pacific Yupik, commonly called Alutiiq, spoken from the Alaska Peninsula eastward to Prince William Sound; Naukanski Siberian Yupik, whose speakers were resettled southward from Cape Dezhnyov, the easternmost point of the Eurasian landmass; Central Siberian Yupik (mainly Chaplinski), which is spoken in the Chukchi Peninsula and on St. Lawrence Island, Alaska; and the very divergent Sirenikski, now virtually extinct.

Aleut. The Aleut language survives in two mutually intelligible dialects: Eastern Aleut, spoken mostly by middle-aged and older people living in eight villages from the Alaska Peninsula westward through Umnak Island, Aleutian Islands, and in the Pribilof Islands in the Bering Sea, which were settled beginning in 1800; and Atkan Aleut, which is spoken also by young people (but no children) on Atka Island, Aleutian Islands, and by some old people on Bering Island, Komandor Islands, Russia, settled in 1826.

Creolized Aleut dialect

Attu, once the westernmost Aleut dialect in Alaska, is now extinct in Alaska, but Attuan Aleut survives on Bering Island in a creolized form (Russian Aleut), with Russian verbal inflections.

The table below indicates the state of the Eskimo-Aleut languages in the last decade of the 20th century.

Table 60: Eskimo-Aleut languages

language	fluent speakers*	population*
Eskimo		
Inuit		
Greenlandic Inuit (Kalaallisut)	46,000	46,400
Eastern Canadian Inuit (Inuktitut)	12,400	14,000
Western Canadian Inuit (Inuktitun)	4,000	7,300
North Alaskan Inuit (Inupiaq)	3,000	15,500
Yupik		
Central Alaskan Yupik	10,000	22,000
Alutiiq Alaskan Yupik	450	2,900
Nankauski Siberian Yupik	50	400
Central Siberian Yupik		
(in Russia)	300	900
(in Alaska, U.S.)	1,050	1,100
Aleut		
Eastern Aleut	110	1,530
Atkan Aleut	45	75
Russian Aleut	10	300

*Statistics gathered 1990–95. Figures do not include residents of urban centres—*e.g.*, about 5,000 Greenlanders in Copenhagen, Denmark.

LINGUISTIC CHARACTERISTICS

Phonological characteristics. Eskimo and Aleut have relatively simple systems of distinctive sounds. The accent (stress) depends upon the length or the number of the syllables and never has independent value as in English.

Vowels. All the languages have the three vowels usually written *a, i,* and *u,* whose pronunciation is determined by the consonants that follow or precede them. Yupik has an additional short *e,* which sounds like the *e* in *roses* or *taxes.*

Consonants. Of distinctive consonants, the Eskimo languages have from 13 to 27, depending on the dialect. The stop sounds include the labial *p,* the dental *t* (made with the tip of the tongue touching the back of the upper teeth), the velar *k,* and the uvular *q* (made with stoppage of the airstream by contact of the back of the tongue and the uvula or back velum); in Alaskan Yupik there is also a palatal *c* (like English *ch*), to which an *s* corresponds in the other dialects. In parts of Canada this has changed to *h.* The nasal sounds, made with the breath passing through the nose, include *m, n,* and *ŋ* (as in "sing"), and, in Yupik, also voiceless nasals (*i.e.,* nasal sounds made without vocal cord vibration). Voiced and voiceless varieties of the continuant consonants *v, l, g,* and the uvular *r*—which is written in Inupiaq and Siberian Yupik with a modified *g*—are distinctive sounds in the western dialects but in eastern Inuit they are only variants. In addition to *y* (written in Canada and Greenland as *j*), some dialects have sounds similar to English *r* or *z* or to *sh* (in Greenlandic written *s*). Corresponding to these, Aleut has a fricative *d* (pronounced as the *th* in *that*). Aleut shares with Eskimo most of the consonants articulated with the tongue, but it has *p* and labial fricatives (*f* and *v*) only in loanwords from Russian or English. Aleut *m* corresponds with Eskimo *m* and *v*; to Eskimo *p* corresponds the Aleut *h* (in initial position) and the Aleut aspirated nasal sound *hm* (pronounced with an accompanying puff of air).

In initial position, Eskimo uses only a single consonant; between vowels at most two. In contrast, Aleut has initial consonant clusters, resulting from the loss of a vowel in the first syllable from an older historical form—*e.g.,* Aleut *sla-* "weather," Inuit *sila.*

The importance of suffixes

Grammatical characteristics. Eskimo has a great number of suffixes but only one prefix and no compounds. In Aleut the word forms are simpler, but syntax can be more complex. Suffixes often are accompanied by changes in the stem, such as the doubling of consonants in Inuit—*e.g., nanuq* "polar bear," dual *nannuk* "two polar bears," plural *nannut* "several polar bears."

Grammatical numbers—singular, dual, plural—combine with suffixes for person—*e.g.,* Inupiaq *ulu-ga* "my knife," and *ulu-t-ka* "my knives," in which *-t-* means "several" and *-ga* or *-ka* "my." The possessor of someone or something occurs in the so-called relative case—*e.g., umialgum pania* "the chief's (his) daughter," in which *pani-a* means "his or her daughter"; this is distinguished from the reflexive *panni* "his or her own daughter" and the stem *panik.*

Yupik and Inuit nouns and pronouns have, respectively, five and six adverbial cases, expressing relations such as "in," "to," "from," "along," "with," and "like"—*e.g., iglu-mi* "in the house," *iglu-ptiŋ-ni* "in our house." Adverbial cases are reduced to two (in/to, from/along) in Aleut and are limited to pronouns and relation words—*e.g., ula-m nag-a-n* "of house (*ula-m*) in its (*-a-*) interior"; this corresponds to Inupiaq *iglu-m ilu-a-ni* "in the house."

In Eskimo, an adverbial case is used also to mark an indefinite object—*e.g.,* singular *-mek* in Alaskan Yupik: *arnaq neq-mek ner'uq* (*r'* = *rr*) "the woman is eating (a) fish," where the subject is in the absolutive case (comparable with a nominative) and the verb in the simple singular (no suffix). A definite object, on the contrary, is in the absolutive case, while the subject is in the relative case (also used as a genitive) and the verb has a suffix referring both to the object and to the subject: *arna-m neqa nera-a* "the woman is eating the fish," *nera-a* "she (or he) is eating it." In Aleut, where the nouns have lost the oblique cases, the system has been transformed into a distinction such as between a specified object in the absolutive case, with a nominal subject in the absolutive case as well, and an anaphoric object, with a nominal subject in the relative case: *ayaga-x̂ qa-x̂ qaku-x̂* "the woman is eating (a or the) fish" (*-x̂* is a singular suffix) versus *ayaga-m qaku-u* "the woman is eating it." An object in the first or second person is marked by a final suffix in Eskimo and by an independent pronoun in Aleut—*e.g.,* Yupik *ikayura-a-nga,* Aleut *ting kiduku-x̂* "he is helping me."

Verbal modes include indicative ("he goes"), interrogative ("did he go?"), imperative ("go!"), optative ("may he go"), participles ("going, gone"), and other forms corresponding to English-language subordinate clauses—*i.e.,* clauses beginning with "if," "when," and so forth. Other modal relations and tenses are specified by derivational suffixes and, in Aleut, by auxiliary verbs—*e.g., haqa-l saĝa-nax̂* "coming he slept" is equivalent to "he came yesterday." An Eskimo derivative form may also correspond to an English complex sentence—*e.g.,* Inupiaq *tikit*$_1$*-qaaġ*$_2$*-mina*$_3$*-it*$_4$*-ni*$_5$*-ga*$_6$*-a*$_{7-8}$ is "he (A)$_8$ said$_{5-6}$ that he (B)$_7$ would not$_4$ be able$_3$ to arrive$_1$ first$_2$," or, in exact Eskimo order, "to arrive first be able would not said him he."

Vocabulary. A remarkable feature of the vocabulary is the great number of demonstratives, about 30 in Inupiaq and Yupik and in Aleut. For example, in Aleut there is *hakan* "that one high up there" (as a bird in the air), *qakun* "that one in there" (as in another room), and *uman* "this one unseen" (heard, smelled, felt).

Local variance

The vocabulary naturally has its local particularities, the various groups having lived under very different conditions. The word that means "meat" in Inuit and some Yupik languages from Greenland to Siberia means "fish" south of Norton Sound and also in Aleut. Word taboo has also played its part, as in East Greenlandic, in which the usual Eskimo-Aleut word for "eye" has been replaced by *uitsatai* (*ui-sa* + *uta-i*) "those by which he keeps gazing."

Eskimo-Aleut derived words (*i.e.,* words that are formed in the way that such English words as "winter-ize" or "anti-dis-establish-ment-ari-an-ism" are formed) correspond quite often to simple, nonderived English words. The possibility of derivation is virtually unlimited in the languages, while the number of word stems is comparatively small. Examples of derivatives in Greenlandic are: *nalu-voq* "is ignorant," *nalu-vaa* "does not know it," and *nalu-na-ar-* "make not (*-er-*) to be (*-na-, -nar-*) ignored," which is equivalent to "communicate." *Nalunaar-asuar-ta-at* "that by which (*-ut*) one communicates habitually (*-ta-, -tar-*) in a hurry (*-asuar-*)," a term (now little used) coined in the 1880s, is Greenlandic for "telegraph."

Greenlandic contains four loanwords from medieval Norse; from the colonial period after 1721 there have been surprisingly few borrowings until the mid-20th century. In Aleut and in the Yupik of former Russian Alaska there are many borrowings from Russian, and there are several in Siberian Eskimo from English, and many from Chukchi, a Paleo-Siberian language. Notable Eskimo contributions to the vocabulary of English and other European languages are "igloo" (*iglu*) and "kayak" (*qayaq*).

Alphabets and orthography. The first book in an Eskimo language was published in 1742 by Hans Egede, a Dano-Norwegian missionary to Greenland. It was printed in the current Roman alphabet.

In 1851 Samuel Kleinschmidt, a German missionary of the Moravian Brethren, systematized the Greenlandic orthography, introducing a special letter and three accents to represent the distinctive sounds of the language. In 1973 the Kleinschmidt orthography was replaced by an orthography in the current Roman alphabet. Numerous publications have appeared in both orthographies.

Moravian missionaries to Labrador, in Canada, published books in Inuit (there called Inuttut) beginning in the early 19th century and toward the end of the century standardized the orthography according to Kleinschmidt's principles. In 1855 the syllabic characters originally designed for the Ojibwa and Cree Indians were introduced to the Inuit of the eastern Arctic, where they are still in use. The Roman alphabet was introduced at a later date to the Inuit of the western Arctic. In 1976 a systematic orthography in the Roman alphabet was proposed for all the Inuit of Canada. In Alaska, Protestant missionaries beginning in 1948 developed for Alaskan Inuit (Inupiaq) a Roman orthography with seven additional letters (now reduced to six).

For the Alaskan Yupik, Moravian missionaries in the 1920s used the ordinary Roman alphabet. In 1971 and 1972 linguists of the University of Alaska in Fairbanks designed systematic practical orthographies in the Roman alphabet for Alaskan Yupik (with three auxiliary accents) and for Siberian Yupik of St. Lawrence Island, both used in many publications. On the Russian side, a Roman alphabet with two additional letters was introduced for Siberian Yupik in 1932, but in 1937 it was replaced by the Russian Cyrillic alphabet.

An adequate Cyrillic alphabet was designed for the Aleut language by the Russian Orthodox missionary Ivan Veniaminov (now known as St. Innocent) about 1830 and was used in religious translations. In 1972 a new Roman orthography with two additional letters was designed by Knut Bergsland for use in the Aleut schools of Alaska.

(K.B.)

North American Indian languages

The term North American Indian languages usually refers to those languages that are indigenous to the United States and subarctic Canada and that are spoken north of the Mexican border. A number of language groups within this area, however, extend as far south as Central America. The present article focuses on the native languages of Canada and the United States. (For further information on the native languages of Mexico and Central America, see below *Meso-American Indian languages*; for most of the languages of Arctic America, see above *Eskimo-Aleut languages*).

The Indian languages of North America are both numerous and diverse. Their original number has been estimated at 300; these tongues were spoken by a native population of approximately 1.5 million. The number of languages still used was estimated at about 200 by the American linguist Wallace Chafe in 1962. Some of these had only one or two elderly speakers. The numbers continue to drop, but with some notable exceptions; *e.g.,* Navajo is steadily increasing in number of speakers. As a consequence of the growing trend toward extinction in the American Indian languages, the field of study is becoming more concerned with the past than with the future. Even so, the rich diversity of these languages provides a valuable laboratory for linguistic theory; certainly the discipline of linguistics could not have developed as it has, especially in the United States, without the Native American languages. In this article, the present tense will be used in referring to both extinct and surviving languages.

Within the diversity of the North American Indian languages, no general characterization is possible; various features of structure are common to them, but there is no feature or complex of features shared by all. At the same time, there is nothing primitive about these languages. They draw upon the same linguistic resources and display the same regularities and complexities as do the languages of Europe. If historical connections are sought among the Indian tongues, some languages clearly show numerous and systematic resemblances comparable to those between Spanish, French, and Italian. These similarities strongly suggest classification as a linguistic family. North American Indian languages can then be grouped into some 57 families. On this level, too, the diversity of some areas is notable. Thirty-seven families lie west of the Rocky Mountains and 20 in California alone; California thus shows more linguistic variety than all of Europe. Some families seem to be related to each other in more remote historical groupings, often called phyla. Such classifications border on speculation, however, partly because data are lacking on many languages (because they are extinct or still unstudied) and partly because of the difficulty in distinguishing, at the deeper historical levels, between resemblances caused by common origin and those resulting from linguistic borrowing.

Language families and phyla

In any case, no theory of common origin for the North American Indian languages has become established. Although most anthropologists believe that North America was populated mainly by people who migrated from Asia across the Bering Strait land bridge, attempts to relate Native American languages to Asian languages have not gained general acceptance. (There is one possible exception—the relationship of Eskimo-Aleut languages to certain Siberian languages.) The linguistic diversity of native North Americans suggests, indeed, that the area was populated as a result of several waves of migration by peoples of distinct linguistic stocks of Asia; these stocks may have no modern survivors.

Classification. The first comprehensive classification into families of the North American Indian languages was made in 1891 by the American John Wesley Powell, who based his study on impressionistic resemblances in vocabulary. A principle of nomenclature adopted by Powell has been widely used ever since: families are named by adding *-an* to the name of one prominent member; *e.g.,* Caddoan is the family including Caddo and other languages. For this most obvious level of relationship, the Powell classification remains essentially unchallenged. Various scholars, however, have attempted to group the families into larger units that reflect deeper levels of historical relationship. Of these efforts, one of the most ambitious and best-known is that of Edward Sapir, which was first published in the *Encyclopædia Britannica* in 1929. In Sapir's classification, all the languages are grouped into six phyla—Eskimo-Aleut, Algonkian Wakashan, Na-Dené, Penutian, Hokan-Siouan, and Aztec-Tanoan—established on the basis of very general grammatical resemblances. In 1958, the American linguist Mary R. Haas established precise sound correspondences between the Algonkian languages and a "Gulf" group in the southeastern United States that Sapir had assigned to the Hokan-Siouan phylum. Since that time, various reconsiderations of Sapir's groupings have been proposed. A classificatory map published by Charles F. and Florence M. Voegelin in 1966 offers one such classification, and it is likely to serve as a standard reference point for some time. While it preserves Sapir's Eskimo-Aleut, Na-Dené, Penutian, and Aztec-Tanoan groups, it proposes reconstituted Macro-Algonkian, Macro-Siouan, and Hokan phyla and allows nine families to remain unclassified, pending further research.

Sapir's six phyla

Table 60, based on the Voegelin map, gives approximate indications of the aboriginal home territories and of the number of speakers estimated from published data in the early 1980s.

Language contact. The Indian languages of North America, like all languages in the world, have always existed in contact with other tongues. From this situation bilingualism, or multilingualism, has resulted; the extent is determined by sociological factors. The Indian languages show varying degrees of linguistic acculturation; *i.e.,* there may be borrowing between languages not only of vocabulary items but also of phonological, grammatical, and semantic features. In aboriginal times, in areas where bilingualism was most important (*e.g.,* the Northwest), there tended to be well-defined linguistic areas in which languages of diverse genetic affiliations came to share numerous structural characteristics through the process of borrowing. As noted above, such phenomena create difficulties for attempts at genetic classifications. In a few cases, situations of language contact have given rise to a pidgin or compromise language that is composed of elements from various sources and is used as a second language, especially in trading. An example is the Chinook Jargon of the Northwest; this came to be used by many whites and absorbed many loanwords from French and English before its eventual obsolescence.

Pidgin languages

In more recent times, contact of Indian languages with European languages—French, English, Spanish, and Russian—has again resulted in bilingualism. With the Indian languages generally relegated to a socially subordinate position (and with many of them headed for extinction), borrowing, however, has involved the relatively superficial level of vocabulary more often than the deeper levels of language structure, such as the sound system or grammar. The effects on European languages are apparent mainly in place names like Massachusetts and Seattle and in names like squash and abalone for native American plants and animals. Among the Indians, the type and degree of linguistic adaptation to European culture has varied greatly, depending on sociocultural factors. For example, among the Karok of northwestern California, a tribe that suffered harsh treatment at the hands of whites, there are only a few loanwords from English (*e.g.*, *ápus* "apples"), a few calques or loan translations (the "pear" is called *vírusur* "bear," because English "pear" and "bear" are merged in Karok pronunciation), but a large number of new formations from native materials; *e.g.*, a hotel is called *amnaam* "eating place."

Grammar. The term grammatical structure as used here refers to both the traditional categories of morphology—how words are made up—and syntax—how words are combined into sentences. It should again be emphasized that in grammar, as well as in phonological or semantic structure, neither the American Indian languages nor any other languages in the world display anything that could be called primitive in the sense of undeveloped or rudimentary. Every language has a structure as complex, as subtle, and as efficiently adaptable to cultural needs as that of Latin or English, for example.

The North American Indian languages display great diversity, so that it is not possible to characterize them as a group by the presence or absence of any particular grammatical peculiarities. At the same time, there are some characteristics that, though not unknown elsewhere in the world, are sufficiently widespread to be considered typical of the continent or of particular linguistic areas within North America. The phenomenon of polysynthesis, in which many sentence elements are expressed within the boundaries of a single word by compounding and affixation, is especially characteristic of Eskimo and Algonkian, but is also found elsewhere. An illustration from the Algonkian group is the Menominee form *nekeespesteh-wenah-nɛɛwaaw* "but I did see him on the way." Incorporation, the compounding of a noun with a verb, is rarely used in English (*e.g.*, "to baby-sit") but is common in some Indian languages; *e.g.*, Mohawk *ke-wẽna-weiẽhõ* "I-language-understand." (The symbols used that are not found in the Latin alphabet have been adopted from phonetic alphabets.)

Polysynthesis and incorporation

Some especially common characteristics of North American languages are the following:

1. In verbs, the person and number of the subject are commonly marked by prefixes; *e.g.*, Karok has *ni-'áhoo* "I walk," *nu-'áhoo* "he walks." In some languages, the prefix simultaneously indicates the object as well as subject; *e.g.*, Karok *ni-mmah* "I see him," *ná-mmah* "he sees me."

2. Tense and aspect of verbs are usually marked by suffixes, as in many languages throughout the world. But in some areas—*e.g.*, among the Athabascan languages—prefixes are used. For example, Chipewyan *hɛ-tsaɣ* means "he is crying," *ɣĩ-tsaɣ* is "he cried," and *ɣwa-tsaɣ* is "he will cry."

3. In noun forms, the concept of possession is widely expressed by prefixes indicating the person and number of the possessor. Thus Karok has *ávaha* "food," *nani-ávaha* "my food," *mu-ávaha* "his food," etc. When the possessor is a noun, as in "man's food," a construction like *ávansa mu-ávaha* "man his-food" is used. Many languages have inalienable nouns, which cannot occur except in such possessed forms. These generally designate such things as kinsmen or body parts; *e.g.*, Luiseño, a language in Southern California, has *no-yó'* "my mother," *o-yó'* "your mother," but no word for "mother" in isolation.

4. Nouns in many languages have forms with a meaning of location; *e.g.*, Karok *áas* "water," *áas-ak* "in the water." Such a construction is reminiscent of the case forms of Latin, and case systems do indeed occur in California and the southwest. For example, Luiseño has the nominative *kíiča* "house," accusative *kíiš*, dative *kíi-k* "to the house," ablative *kíi-ŋay* "from the house," locative *kíi-ŋa* "in the house," instrumental *kíi-tal* "by means of the house."

The following five grammatical features are less typically North American, but are nevertheless distinctive of many areas. First person pronouns in many languages show a distinction between a form inclusive of the addressee—"we" denoting "you and I"—and an exclusive form—"I and someone other than you." Some languages also have a distinction in number between singular, dual, and plural pronouns. Reduplication, the repetition of all or part of a stem, is widely used to indicate distributed or repeated action of verbs; *e.g.*, in Karok, *imyah* means "breathe," *imyáhyah* means "pant." In Uto-Aztecan languages, reduplication sometimes is associated with plural nouns, as in Pima *gogs* "dog," *go-gogs* "dogs." In many languages, verb stems are distinguished on the basis of the shape or other physical characteristics of the associated noun; thus in Navajo, in referring to motion, *'áⁿ* is used for round objects, *táⁿ* for long objects, *tíⁿ* for living things, *lá* for ropelike objects, etc. Similar distinctions may refer to dual and plural number. Karok has *ikpuh* "one swims," *iθpuh* "two swim," *ihtak* "several swim."

Inclusive and exclusive pronouns

Verb forms also frequently specify the location or direction of an action by the use of prefixes or suffixes. In Karok, for example, from *paθ* "throw" is derived *páaθroov* "throw upriver," *páaθ-raa* "throw uphill," *paaθ-rípaa* "throw across-stream," and as many as 38 other similar forms. Some languages also specify the instrument of an action, generally by prefixation; *e.g.*, Pomo *phi-de-* "to move by batting with a stick," *phu-de-* "to move by blowing," *pha-de-* "to move by pushing with the end of a stick." Lastly, many languages have evidential forms of verbs that indicate the type of validity of the information reported; such distinctions may assume the importance played by tense and aspect in European languages. Thus Hopi distinguishes *wari* "he ran, runs, is running" as a reported event, from *warikŋwe* "he runs (*e.g.*, on the track team)," which is a statement of general truth, and from *warikni* "he will run," which is an anticipated event. In other languages verb forms consistently discriminate hearsay from eye-witness reports. Such a system might be very welcome in other societies; *e.g.*, especially as regards the reliability of news reports.

Phonology. The languages of North America are as diverse in their systems of pronunciation as they are in other ways. In terms of the number of contrasting sounds (phonemes), the Northwest Coast is characterized as a linguistic area by the unusual richness of its systems. A language like Tlingit has approximately 50 consonants and vowels (a comparable count for English would number 35). By contrast, Karok has only 23. The richest sound inventories seem to occur where bilingualism was commonest, and sounds were borrowed between languages.

The large number of consonants that is found in many Indian languages is based on the use of a number of phonetic contrasts that are relatively unfamiliar in European languages. In English, different consonants are produced by vibrating the vocal cords (which results in voiced sounds) or by not vibrating them (which gives unvoiced sounds); by shutting off the air momentarily, thus producing stops, or by letting the airstream pass through the mouth with friction (producing fricatives); and by placing the tongue in a variety of positions. The Indian languages also use

Consonant features

Table 61: North American Indian Languages*

phyla, families, languages	location	speakers remaining†
American Arctic-Paleosiberian		
Eskimo-Aleut		
Chukchi-Kamchatkan	Siberia	
Na-Dené		
Athabascan		
Dogrib, Bear Lake Hare	N.W.T.	1,400
Chipewyan, Slave, Yellowknife	N.W.T.	4,400+
Kutchin	Yukon, Alaska	800
Tanana, Koyukon, Han, Tutchone	Alaska	1,450+
Sekani, Beaver, Sarsi	Alberta	450+
Carrier, Chilcotin	B.C.	1,500+
Tahltan, Kaska	N.W.T.	300+
Tanaina, Ingalik, Nabesna, Ahtena	Alaska	1,500+
Eyak	SC Alaska	‡
Chasta Costa, Galice, Tututni	SW Oregon	‡
Hupa	NW Calif.	1,200
Kato, Wailaki	NC Calif.	235
Mattole	NC Calif.	§
Tolowa	NW Calif.	125
Navajo	Ariz., N.M.	137,400+
Western Apache	W Arizona	10,000+
Chiricahua, Mescalero Apache	S New Mexico	2,900+
Jicarilla Apache	N New Mexico	1,000
Lipan Apache	Texas	‡
Kiowa Apache	Oklahoma	‡
Tlingit	SE Alaska	1,500+
Haida	B.C.	700
Macro-Algonkian		
Algonkian		
Cree, Naskapi, Montagnais	E Canada	89,200
Menominee	Great Lakes area	2,200+
Fox-Sauk-Kickapoo	south of Great Lakes	1,000
Shawnee	SC U.S.	2,500+
Potawatomi	Michigan	3,500+
Ojibwa, Ottawa, Algonkin Salteaux	S Ontario	25,000+
Delaware	C Atlantic coast	3,200+
Penobscot, Abnaki	New England	300+
Malecite, Passamaquoddy	New England, Maritime Provinces	1,200+
Micmac	Maritime Provinces	2,100+
Blackfoot	Mont., Alberta	1,000+
Cheyenne	E Wyoming	4,000+
Arapaho, Atsina, Nawathinehena	E Colorado	1,000+
Yurok	NW Calif.	1,900+
Wiyot	NW Calif.	1
Muskogean		
Choctaw, Chickasaw	N Mississippi	18,000+
Alabama, Koasati	Alabama	700+
Mikasuki, Hitchiti	NW Florida	1,000
Muskogee (Creek), Seminole	Georgia	8,500+
Natchez	N Louisiana	§
Atakapa	SW Louisiana, SE Texas	§
Chitimacha	S Louisiana	§
Tunica	N Louisiana	§
Tonkawa	E Texas	§
Macro-Siouan		
Siouan		
Crow	E Montana	4,600
Hidatsa	North Dakota	1,000+
Winnebago	Wisconsin	1,000+
Omaha, Osage, Ponca, Kansa, Quapaw	central plains	3,400+
Dakota (Sioux)	northern plains	30,000+
Tutelo, Ofo, Biloxi	Gulf Coast	§

phyla, families, languages	location	speakers remaining†
Catawba	Carolinas	§
Iroquoian		
Seneca, Cayuga, Onandaga	New York	6,800+
Mohawk	New York	6,700+
Oneida	New York	6,200+
Wyandot (Huron)	SE Ontario	§
Tuscarora	North Carolina	600+
Cherokee	southern Appalachians, Oklahoma	27,000
Caddoan		
Caddo	Arkansas	1,200+
Wichita	Texas, Okla.	500+
Pawnee	Kansas	1,800+
Arikara	Dakotas	300+
Yuchi	southern Appalachians	200+
Hokan		
Yuman		
Walapai, Havasupai, Yavapai	NW Arizona	600+
Mohave, Yuma	lower Colorado River	2,000
Delta Yuman (Cocopa)	delta of Colorado River	300+
Diegueño, Kiliwa	S Calif., Baja Calif.	75+
Seri	Sonora	200
Pomo	NC Calif.	
Northern Pomo		40
Northeast Pomo		1
Central Pomo		40
Southwest Pomo		50
Southeast Pomo		10
Southern Pomo		10
Palaihnihan	NE Calif.	
Achomawi		40+
Atsugewi		‡
Shastan	NE Calif.	‡
Yanan	NC Calif.	§
Chimariko	NW Calif.	§
Washo	EC Calif., Nevada	1,100
Salinan	WC Calif.	360
Karok	NW Calif.	750+
Chumashan	S Calif.	360
Comecrudan	S Texas, NW Mexico	§
Coahuiltecan	S Texas, NW Mexico	§
Esselen	WC Calif	§
Branches in Meso-America		
Penutian		
Yokutsan	SC Calif.	800+
Maiduan	NC Calif.	2,500+
Wintun	NC Calif.	
Patwin		50+
Wintu, Nomlaki		1,200+
Miwok-Costanoan		
Miwok (Sierra, Coast-Lake)	SC Calif., C Calif.	2,500
Costanoan	WC Calif.	§
Klamath-Modoc	SC Oregon	2,500+
Sahaptian		
Sahaptin (Klikitat, Umatilla, Walla Walla, Warm Springs, Yakima)	NC Oregon	1,400+
Nez Percé	WC Idaho	500+
Cayuse	NE Oregon	§
Molale	NC Oregon	§
Coos	SW Oregon	50
Yakonan	WC Oregon	
Alsea		§
Siuslaw, Lower Umpqua		§
Takelma	SW Oregon	§
Kalapuya	WC Oregon	‡
Chinookan	NW Oregon, SW Wash.	20
Tsimshian	WC B.C.	3,000
Zuni	WC New Mexico	3,000+
Latin American branches		

phyla, families, languages	location	speakers remaining†
Aztec-Tanoan		
Kiowa-Tanoan		
Tiwa	NC New Mexico	5,300+
Tewa	NC New Mexico	2,400
Towa	NC New Mexico	2,000
Kiowa	Oklahoma	1,000
Uto-Aztecan		
Mono	EC Calif.	300+
Northern Paiute (Paviotso), Bannock, Snake	NE Calif., SE Ore., N Nev., S Idaho	3,500
Panamint, Gosiute, Shoshone	C Nev., N Utah, SW Wyo.	5,000+
Comanche	N Texas	800
Kawaiisu, Ute, Chemehuevi, Southern Paiute	SE Calif., S Nevada, S Utah, SW Colo.	3,000+
Hopi	N Arizona	7,900+
Tubatulabal	SC Calif.	‡
Luiseño	S Calif.	500+
Cahuilla	S Calif.	800+
Cupeño	S Calif.	‡
Serrano	S Calif.	‡
Pima-Papago	S Arizona	25,400+
Latin American branches		
Unclassified		
Keresan	New Mexico	15,800
Yukian	NC Calif.	
Yuki		‡
Wappo		‡
Beothuk	Newfoundland	§
Kutenai	Mont., Idaho, B.C.	600+
Karankawa	SE Texas	§
Chimakuan	NW Washington	
Quileute		200+
Chemakum		§
Salish		
Lilloet	C B.C.	1,000+
Shuswap	E B.C.	1,000+
Thompson	C B.C.	1,000+
Okanagon, Sanpoil, Lake, Colville	S B.C.	1,700+
Pend d'Oreille, Flathead, Spokan, Kalispel	N Idaho	600+
Coeur d'Alene	N Idaho	400
Middle Columbia, Wenatchee	E Washington	30
Tillamook	NW Oregon	100
Twana	NW Washington	1,000
Upper Chehalis, Cowlitz, Lower Chehalis, Quinault	W Washington	1,700+
Southern Puget Sound Salish	W Washington	50+
Straits Salish	SW B.C.	2,300
Halkomelem	SW B.C.	1,000+
Squamish	SW B.C.	100+
Comox, Sishistl	Vancouver Is.	‡
Bella Coola	WC B.C.	200+
Wakashan		
Nootka	Vancouver Is.	1,000+
Nitinat	Vancouver Is.	10+
Makah	NW Washington	1,500
Kwakiutl	WC B.C.	1,000
Bella Bella, Heiltsuk	WC B.C.	100+
Kitamat, Haisla	WC B.C.	100+
Timucua	Florida	§

*Phyla given in boldface type; families given in italics (including those consisting of single languages); single languages, or dialect groups so closely related that they can be treated as single languages, given in roman type. †1981 estimate. ‡Minimal number of speakers; *i.e.*, under 10. §Extinct.

Figure 34: Distribution of North American Indian languages.
From C.F. and F.M. Voegelin, *Map of North American Indian Languages;* copyright 1966 by University of Washington Press

these mechanisms, but sometimes others as well. The glottal stop, an interruption of breath produced by closing the vocal cords (as in the middle of English *oh-oh!*) is a common consonant. A related phenomenon, widespread in western North America, is the use of glottalized consonants, as when a *t* is produced with near simultaneous closure and reopening of the vocal cords. This is recorded with an apostrophe; it differentiates terms like Hupa (Athabascan) *teew* "underwater" from *t'eew* "raw."

The number of consonantal contrasts is also frequently expanded by distinguishing a larger number of tongue positions than do most European languages. Many languages distinguish two types of velar sounds (sounds made with the back of the tongue)—a *k* much like an English *k*, and a uvular *q*, produced further back in the mouth. Some languages even differentiate three such *k* sounds—front, middle, and back. Labiovelars, velar sounds that have simultaneous lip-rounding, are also common. Thus Tlingit has 21 phonemes made in the velar area alone: *g*, *k*, uvular *G*, *q*, glottalized *k'*, *q'*, labiovelar g^w, k^w, k'^w, G^w, q^w, $q^{w'}$, in addition to the corresponding fricatives y and *x*, with uvular *X*, glottalized *x'*, *X'*, and labiovelar x^w, X^w, x'^w, X'^w. In comparison, English has only two sounds, *k* and *g*, made in the same area of the mouth.

Another class of sounds common in North America, especially in the West, is that of the laterals, which are produced by stopping the breath with the central part of the tongue but allowing it to escape at the sides. Alongside the common lateral *l*, such as exists in English, many Indian languages have a voiceless counterpart, similar to the Welsh *ll;* this sound is approximated by the *thl* in northwestern place names such as Cathlamet. To this some languages also add glottalized varieties, as well as a close-knit *tl* unit, which may in turn be aspirated or glottalized, so that there may result, as in Navajo, a total of five distinguishable lateral sounds.

In some Indian languages, as in English, stress is significant in distinguishing the meaning of words. In others, musical pitch plays a linguistic function, as it does in Chinese; *e.g.*, in Navajo, *bíní'* is "his nostril," *bìnì'* is "his face," and *bìní'* is "his waist." (High and low pitches are indicated with the acute and grave accents, respectively.)

A peculiarity of some northwest coast languages is their use of complex consonant clusters, as in Bella Coola *tlkʷixʷ* "don't swallow it." Some words even lack vowels entirely; *e.g.*, *nmnmk'* "animal."

Phonological change

Processes of phonological change, in which differences of sound are associated with grammatical distinctions (as with English *f* and *v* in "half," "halves," "to halve"), are also found in North American languages. In some languages, for example, consonantal change is related to diminutive meaning: thus Luiseño *r* changes to *đ* in *ŋarúŋru-š* "pot," *ŋadúŋdu-mal* "pot-small." Vowel harmony, a process whereby vowels change to resemble adjacent ones, is further attested in North America. Yurok in northwestern California, for example, has an unusual *r* vowel, comparable to the sound in English "bird"; when this occurs in a suffix, stem vowels change to agree with it, thus *lo'oy̥e* "black" + *-'r'y* (animate suffix) yields *lr'r̥yr'r'y* "black animal."

Vocabulary. The word stock of American Indian languages, like those of other languages, is composed both of simple stems and of derived constructions; the derivational processes commonly include affixation (the use of prefixes, suffixes, etc.) in addition to compounding in some languages. A few languages use internal sound change, similar to the case of English "song" from "sing"; *e.g.*, Yurok *pontet* "ashes," *prncrc* "dust," *prncrh* "to be gray." New vocabulary items are also acquired by borrowing, as mentioned above.

It should be noted that, in languages generally, the meaning of a vocabulary item cannot be adequately inferred from a knowledge of its historical origin or from knowing the meaning of its parts. For example, the name of an early 19th-century trapper, McKay, entered Karok as *mákkay*, but with the extended meaning of "white man." It was then compounded with a native noun *váas* "deerskin blanket" to give the neologism *makáy-vaas* "cloth"; this in turn was compounded with *yukúkku* "moccasin" to give *makayvas-yukúkku* "tennis shoes." At each stage of vocabulary formation, meaning is determined not simply by etymology but also by arbitrary extensions or limitations of semantic value.

Semantic structure of Indian languages

It is in the area of semantic structure that American Indian vocabulary is likely to present some surprises to the investigator. It is frequently observed that the immense diversity of the physical universe is reduced by every society to a manageable set of classifications embodied in its vocabulary. But there are few universals in such classification, and every language makes its unique semantic divisions. One language may make many specific discriminations in a particular area, while another is content with a few general terms; the difference is correlated with the importance of the semantic area for the particular society. Thus English is highly specific in classifying bovines (bull, cow, calf, heifer, steer, ox), even to the point of lacking a general cover term in the singular (what is the singular of cattle?), but for other species it has only cover terms like camel, llama. North American Indian vocabularies, as would be expected, embody semantic classifications that reflect native American environmental conditions and cultural traditions.

Interest in the semantic classifications of American Indian languages, especially in Hopi, has been particularly stimulated by the work of the American investigator Benjamin Lee Whorf. When English discriminates "air-plane," "aviator," and "flying insect," Hopi generalizes with a single term *masa'ytaka*, roughly "flier"; but when English uses a single general term, "water," Hopi differentiates *pāhe* "water in nature" from *kēyi* "water in a container."

The vocabularies of different languages may differ not only in the categorization of particular items but also in the general principles of semantic organization; such differences may be found even between neighbouring languages in a single culture area. English, for example, tends to exhaust the universe of flora and fauna with multilevelled hierarchical classifications such as "plant, bush, berry bush, gooseberry bush" or "animal, insect, louse, body louse," but the languages of northwestern California, by contrast, have relatively few generic terms and many vocabulary items that do not fall into any such hierarchy. The generic terms of Yurok refer, roughly, to "quadruped mammal," "fish," "snake," "bird," "tree," "bush," "grass," "flower," and "berry"; the organization in the neighbouring Tolowa language is simpler, lacking "quadruped mammal" and "fish." In such frameworks, a term like Yurok *wrr̥yr* "body louse" cannot be subsumed in the larger classes of "louse" or "insect" because none exist. The placing of terms in semantic pigeonholes tends to be replaced, in these semantic systems, by identifying them in terms of similarity. A Yurok speaker, asked to identify a flowering bush for which he knows no name will describe it not as "a kind of bush," but as *sahsip sey̥on* "similar to wild lilac." Such evidence suggests that the semantic structures of some American Indian vocabularies are based on classes defined less by their boundaries than by their centres.

Kinship terms

Another type of semantic structuring is illustrated by certain systems of kinship terms. In Fox, an Algonkian language, the term for maternal uncle also includes maternal grandmother's sister's son's son (a kind of second cousin). This can be accounted for by recognizing some very simple rules, rules that apply to the other terms of the kinship system as well: (1) siblings of the same sex, as linking relatives, are reckoned as equivalent; (2) a father's sister, as a linking relative, is equivalent to a sister, and conversely, a mother's brother's child is equivalent to a mother's brother. Then a mother's mother's sister's son's son, by rule 1, is equivalent to a mother's mother's son's son; but because one's mother's son is one's brother, this is the same as a mother's brother's son; and this in turn, by the converse of rule 2, is equivalent to a mother's brother. It is clear that the semantic systems of American Indian languages exhibit not only structures of hierarchy and similarity but also rules of semantic equivalence.

Language and culture. The exotic character of American Indian semantic structures, as manifested not only in their vocabularies but also in the relationships expressed by their morphological categories and syntactic patterns, has led a number of scholars to speculate on the relationships between language, culture, and habitual thought patterns or "world view." It was hypothesized that the unique organization of the universe that is embodied in each language might act as a determining factor in the individual's habits of perception and of thought, thus forming and maintaining particular tendencies in the associated nonlinguistic culture. As Edward Sapir put it,

> Human beings do not live in the objective world alone, . . . but are very much at the mercy of the particular language which has become the medium of expression for their society . . . The fact of the matter is that the "real world" is to a large extent unconsciously built up on the language habits of the group . . . We see and hear and otherwise experience very largely as we do because the language habits of our community predispose certain choices of interpretation.

The Whorfian hypothesis

This idea was further developed, largely on the basis of work with American Indian languages, by Sapir's student Benjamin Lee Whorf, and is now often known as the Whorfian hypothesis. Whorf's initial arguments focussed on the strikingly different organization of experience that can be found between English and Indian ways of saying "the same thing." From such linguistic differences, Whorf infers underlying differences in habits of thought. It then remains to show how these habits are manifested in nonlinguistic cultural behaviour. Thus, Whorf points out that, in Hopi, words referring to units of time (*e.g.*, "day") differ from other nouns in that they have no plural form; furthermore, they cannot be counted with the cardinal numerals ("one," "two," etc.) but only with the ordinals ("first," "second," etc.). From this he infers that when the English speaker speaks of "ten days," as if the days were an aggregate of separate units, the Hopi speaker, on the other hand, thinks in terms of the cyclic recurrence of a single phenomenon. Whorf attempts to support this idea by reference to Hopi ceremonial behaviour, which involves repeated preparation for future events. If, in the Hopi view, each day is really a recurrence, rather than something new, then it is reasonable to believe that the daily repetition of ceremonial acts will have a cumulative effect on the future. As Whorf says, the Hopi belief is diametrically opposed to the English proverb that "Tomorrow is another day."

More investigation is necessary to either prove or disprove the Whorfian hypothesis. In any case, the diversity of American Indian languages and cultures has continued to provide a rich laboratory for investigation. A particularly interesting problem is found in the area of northwestern California, where several small tribes have very similar cultures, but use languages of very diverse types. These are Karok, genetically classified as Hokan; Yurok and Wiyot, which are Algonkian; and Hupa and Tolowa, Athabascan languages. By the Whorfian hypothesis, one might expect that the difference in languages would have produced a greater diversity in the cultures; or failing that, one might expect the languages to have grown more similar to each other. In fact, both linguistic diversity and cultural uniformity seem to have made modest accommodations to each other. As an example of Whorfian linguistic determinism, the systems of biological taxonomy of Yurok and Tolowa, referred to in the previous section, may be noted. The Yurok have a larger number of generic classifications, which means they have more choice in nomenclature, because either a generic or a specific term can be used. This is consistent with the high degree of choice afforded in Yurok grammar, in which word order is nearly free and many morphological categories are optional. The sparser taxonomy of Tolowa offers less choice, corresponding to a much more rigid grammatical structure.

Language, culture, and prehistory

A different kind of relationship between language and culture is of more interest to the student of North American prehistory, namely, the fact that language retains traces of historical changes in culture and so aids in reconstructing the remote past. Here again the pioneering work was done by Sapir, who pointed out, for instance, that the original home from which a group of related languages or dialects has dispersed is more likely to be found in the area of great linguistic diversity; *e.g.*, there are much greater differences in the English dialects of the British Isles than of the more recently settled areas such as North America or Australia. To take an American Indian example, the Athabascan languages are now found in the Southwest (Navajo, Apache), on the Pacific Coast (Tolowa, Hupa), and in the Western Subarctic. The greater diversity of the Subarctic languages leads to the hypothesis that the original centre of Athabascan migration was from that area. This northern origin of the Athabascans was further confirmed in a classic study by Sapir in which he reconstructed parts of prehistoric Athabascan vocabulary, showing, for example, how a word for "horn" had come to mean "spoon" as the ancestors of the Navajo migrated from the far north (where they made spoons of deerhorns) into the Southwest (where they made spoons out of gourds). The correlation of such linguistic findings with the data of archaeology holds great promise for the study of American Indian prehistory.

Writing and texts. Although a writing system was in use among the Mayas of Meso-America at the time of first European contact, none was known in North America. All writing systems that have been used for North American Indian languages have resulted from the stimulus of European writing, or have actually been invented and introduced by whites. Perhaps the most famous system is that invented by Sequoyah, a Cherokee, for his native language. It is not an alphabet but a syllabary, in which each symbol typically stands for a consonant-vowel sequence. The forms of characters were derived in part from the English writing system, but without regard to their English pronunciation. Well suited to the language, the syllabary fostered widespread literacy among the Cherokee until their society was disrupted by government action; its use, however, has never died out, and attempts are now being made to revive it.

Other writing systems, invented by missionaries, teachers, and linguists, have also included syllabaries; *e.g.*, for Cree, Winnebago, and some northern Athabascan languages. Elsewhere, alphabetic scripts have been used, adapted from the Roman alphabet by the use of additional letters and diacritics. White educational policy, however, has generally not encouraged literacy in Indian languages. A rich oral literature of American Indian myths, tales, and song texts has been in part published by linguists and anthropologists, and there is now increasing encouragement for the training of Indians to transcribe their own traditions—*e.g.*, among the Navajo. It is possible that there may yet be a flowering of American Indian literature, not only in spoken but also in written form. (W.O.B.)

Meso-American Indian languages

Meso-American, or Middle American, Indian languages are spoken in an area of the aboriginal New World that includes central and southern Mexico, Guatemala, Belize (British Honduras), El Salvador, parts of Honduras and Nicaragua, and part of northwest Mexico. Though various centres of civilization have flourished in the area, sometimes concurrently, from 1000 BC down to the time of the Spanish conquest of Mexico in 1519, Meso-America as a whole has had a more or less common cultural history for 2,500 years.

Treatments of the languages of Meso-America are customarily organized on the basis of their genetic relationships, and only secondarily on that of geographical distribution. Thus, some languages treated as Meso-American are not in fact spoken in Meso-America proper but form linguistic families with languages that are spoken there. For information about languages of northeast, north central, and northwest Mexico that are not dealt with in this section, see above *North American Indian languages*. For languages of Central America not treated here, see below *South American Indian languages*.

Numbers of languages and speakers

Some 70 Indian languages are spoken today in Meso-America by perhaps 7,500,000 people. When the Spanish conquered Mexico in 1519, there may have been 20,000,000 people in Meso-America. Within 100 years of the conquest, the Indian population had decreased by 80 percent as a result of war, disease, forced labour, and starvation. Since then the Indian population has gone back to a higher level, but several languages—have become extinct. Meso-American languages with the greatest number of speakers in the mid-20th century are:

language	number of speakers	family
Aztec	1,200,000	Uto-Aztecan
Yucatec	600,000	
Quiché-Tzutujil-Cakchiquei	1,200,000	Mayan
Mam	450,000	
Kekchi	375,000	
Mixtec	350,000	
Zapotec	400,000	Oto-Manguean
Otomi	450,000	

The study of the Meso-American languages. During the 16th and 17th centuries, some Dominican and Franciscan missionaries devoted themselves to the study of native languages so that priests could deal in religious matters with monolingual Indians. They wrote grammars following a Latin model, devised orthographies applying values used in Spanish or Latin (occasionally inventing new letters), made dictionaries (usually vocabularies or glossaries), and translated Christian texts (confessionals, sacraments, and sermons) into Indian languages. Except for one heroic figure, the Spanish missionary priest Bernardino de Sahagùn, they neither collected nor fostered the collection of folklore. During this period grammars and dictionaries were written for such languages as Aztec, Zapotec, Mixtec, Tzeltal, Yucatec, Quiché-Tzutujil-Cakchiquel, Chortí, and Northwestern Otomí. These collections of data served the successors of the first missionaries. During the 18th century, the momentum of such work decreased, and,

Table 62: Meso-American Indian Languages

family, branch (or group), language	location	number of speakers
1. Uto-Aztecan (Uto-Nahuan) family 48c*		
Shoshonean (Yutan, Otegonian) division† 34c		
A. *Plateau group 18c*		
1. Mono, N Paiute—Bannock (complex?)		
2. Shoshoni—Gosiute (Goshiute), Comanche (complex?)		
3. Ute-Chemehuevi, S Paiute (complex?)		
B. Tubatulabal		
C. *Southern California branch 24c*		
1. Serrano		
2. Luiseño, Juaneño		
3. *Gabrieleño complex 10c*		
Gabrieleño		
Fernandeño		
4. *Cahuilla complex‡*		
Cahuilla		
Cupeño		
D. Hopi		
Sonoran (Mexican) division 39c		
E. *Piman group*		
1. *Piman complex 8c*		
Papago (Pima)	Arizona; Sonora	500
Lower Pima (Nevome)	Sonora	900
Tepecano§	Jalisco	a few
2. *Tepehuán complex‡*		
N Tepehuán§	Sonora	6,300
S Tepehuán§	Jalisco	17,700
F. *Yaquian (Taracahitian) branch 23c*		
1. *Tarahumara complex 7c*		
Tarahumara (Rarámuri)	Chihuahua	36,600
Guarillo		7,200
2. Tubar		extinct
3. *Cáhita complex 15c*		
Eudeve (Heve)		extinct
Ópata, Jova		extinct
Yaqui, Mayo (Cáhita)§	Arizona; Sonora, Sinaloa	26,500
G. *Coran group 15c*		
Cora§		?
Huichol§	Nayarit	10,900
H. *Nahuan group 15c*		
1. *Aztec complex 11c*		
C, N Aztec (Nahuatl)§	State of Mexico, Puebla, Hidalgo	1,200,000 (C, N, W and E Aztec together)
W Aztec (Nahual)§	Michoacán	
E Aztec (Nahuat)§	Veracruz	
Pipil§	C America	2,000
2. Pochutec§	Oaxaca coast	extinct
2. Cuitlatec (Teco) isolate	Guerrero	extinct
3. Seri isolate	Sonora coast	400
4. Tequistlatec complex or group	SE Oaxaca	
Huamelultec	coastal region	5,000
Tequistlatec	mountain region	5,000
5. Jicaque isolate (several dialects or languages)	NW Honduras	300
6. Tlapanec (Subtiaban, Tlapanecan) complex 8c		
A. Tlapanec (Yope)	Guerrero	44,300
B. Subtiaba (Nagrandan)‖	Nicaragua	extinct?
Maribio‖	El Salvador	extinct?
7. Oto-Pamean stock 55c		
A. Chichimec (Meco, Jonaz)	Guanajuato	1,000
B. *Pamean group 18c*		
N Pame	San Luis Potosí	3,600
S Pame	Hidalgo	
C. *Matlatzinca complex 10c*	State of Mexico	
Matlatzinca (Pirinda)		2,800
Ocuiltec (Atzingo)		a few
D. *Otomian group 16c*		
1. *Otomi complex 9c*	Hidalgo, Guanajuato, State of Mexico, Querétaro	432,000
NW Otomi		
NE Otomi		
SW Otomi		
Ixtenco Otomi		
2. Mazahua	Michoacán, State of Mexico	221,000
8. Popolocan (Mazatecan) family 25c		
A. *Chochoan group 13c*		
1. Ixcatec	NW Oaxaca (Santa Maria Ixcatian)	200
2. *Chocho complex 8c*		
Popoloc	SE Puebla, NW Oaxaca	34,000
Chocho	NW Oaxaca	2,500
B. *Mazatec complex 10c*		145,500
Mazatec (1)	SE Puebla, N Oaxaca	
Mazatec (2)		
9. Mixtecan family 42c		
A. Amuzgo	E Guerrero, W Oaxaca	20,100
B. *Greater Mixtecan branch 25c*		
1. *Mixtec group 15c*	E Guerrero, S Puebla, W Oaxaca	335,100
Mixtec (1)		
Mixtec (2)		
Mixtec (3)		
2. Cuicatec	NE Oaxaca	20,200
C. Trique	W Oaxaca	18,700
10. Zapotecan family 24c	Oaxaca	
A. *Zapotec group 14c*		407,600
Juárez Zapotec	Ixtlán	
Villalta Zapotec	Yatzachi	
S Mountain Zapotec	Cuixtla	
Valley Zapotec	Mitla, Tehuantepec	
B. Soltec	Sola de Vega	extinct
C. Papabuco	Elotepec	extinct
D. Chatino	southwest	27,500

*Indicates centuries of separation. †Not spoken in Meso-America. ‡There is some doubt whether these groups should be given the status of complexes. §Sonoran languages spoken in Meso-America. ‖Varieties of the same language spoken in different countries (and having different names).

after Mexico became independent in the first part of the 19th century, Spanish clerics were ousted, leaving further work on indigenous languages to travellers and gentlemen scholars—mostly people who were poorly qualified for such a task.

Extent of the studies of Meso-American languages

Modern linguistic techniques for the description of languages were not applied to Meso-American languages until North Americans turned their attention to the area in the 1930s and 1940s. Since then, much professional linguistic work has been done on these languages, especially those of Mexico. Almost every language of Meso-America has been worked on by at least one linguist, but the time spent and level of linguistic competence of the investigators have varied greatly. For most of the languages, grammatical and lexical data have been collected, much of which remains unpublished. A number of competent grammars and dictionaries have appeared; none of them however, is exhaustive or definitive. Folktales have been collected for a smaller number of languages. Spanish-based orthographies have been devised for most of the Meso-American languages in the 20th century, but not much reading matter is available in them. In short, much work remains to be done.

CLASSIFICATION

Modern genetic groupings. The classification of Meso-American Indian languages presented here reflects generally accepted genetic groupings (as of the early 1970s), based on similarities in vocabulary and grammar and on the establishment of regular correspondences between sounds in cognate (related) words among the several languages. The languages grouped together are presumed to have developed from a common ancestor, called a protolanguage. Not all of the languages of Meso-America have been convincingly assigned to a specific group. A few of these languages are currently thought to be unrelated to any of the established genetic groupings and are listed individually in the table; these solitary languages are called isolates.

Within a given genetic grouping of languages, there may be several levels of relatedness. Glottochronology (or lexicostatistics), developed by two linguists in the United States—Morris Swadesh and Robert Lees—is a controversial and not universally accepted procedure for measuring the degrees of difference between related languages in terms of years of separation. Based on the assumption that all languages change more or less to the same degree within a given period of time, the method employs a list of 100 items of "basic" or "noncultural" concepts, which are assumed to be expressible by vocabulary items in any language. Over the millennia, different words will have been substituted to express 14 percent of the 100 concepts every 1,000 years, and two languages that separated 1,000 years ago will share 74 percent cognates (86 percent of 86

Table 62: Meso-American Indian Languages (continued)

family, branch (or group), language	location	number of speakers
11. Chinantecan group 15c*	N Oaxaca	80,000
Chinantec (1)		
Chinantec (2)		
Chinantec (3)		
Chinantec (4)		
12. Manguean (Chorotegan, Chiapanec-Mangue) group 13c		
A. Chiapanec	Chiapas	extinct
B. Mangue (Dirian, Nagrandan)‖	Nicaragua	extinct
Chorotega‖	Honduras	extinct
Nicoya (Orotiña)‖	Costa Rica	extinct
13. Huave isolate	SE Oaxaca	25,300
14. Mixe-Zoque (Zoquean, Mixean, Zoque-Mixe) family 36c		
A. *Zoquean group 14c*		
Zoque	Tabasco, Chiapas, Oaxaca	37,600
Sierra Popoluca	Veracruz	25,300
Texistepec	Veracruz	3,000
B. *Mixean group 13c*		
1. Sayula	Veracruz	1,000
Oluta	Veracruz	1,000
E, W Mixe	E Oaxaca	77,500
2. Tapachultec	SE Chiapas coast	extinct
15. Totonacan family 2bc		
Totonac	Veracruz, Puebla	239,000
Tepehua	Veracruz, Hidalgo	18,800
16. Mayan family 41c		
A. *Huastec complex 9c*		
Huastec	San Luis Potosí, N Veracruz	101,000
Chicomuceltec (Coxoh)	Chiapas	a few?
B. *Yucatec (Maya) complex 10c*		
Yucatec	Yucatán, Campeche, Quintana Roo, N Guatemala, Belize	605,000
Lacandón	Chiapas	200
Itzá	N Guatemala	500
Mopán	N Guatemala, Belize	6,000
C. *Western division 30c*		
1. *Greater Tzeltalan branch 19c*		
a. *Cholan proper group 14c*		
Chontal (Yocotán)	Tabasco	51,000
Chol	Tabasco, Chiapas	109,000
Chorti	Honduras, E Guatemala	64,000
b. *Tzeltalan group 14c*	Chiapas	
Tzotzil (Quelén)		123,000
Tzeltal		123,000
2. *Greater Kanjobalan branch 21c*		
a. *Chujean group 16c*		
Tojolabal (Chaneabal)	Chiapas	19,000
Chuj	NW Guatemala	30,000
b. *Kanjobal proper group 15c*		
i. *Kanjobal complex 7c*	NW Guatemala	
Kanjobal (Conob, Solomec)		62,000
Acatec		13,000
Jacaltec		21,000
ii. *Mochó complex* (Cotoque)	SE Chiapas	
Motozintlec		500
Tuzantec		100
D. *Eastern division 34c*		
1. *Greater Mamean branch 26c*		
a. *Mamean proper group 15c*		
Teco	SE Chiapas, W Guatemala	5,000
Mam	W Guatemala	434,000
b. *Ixilan group 14c*	NW Guatemala	
Aguacatec		15,000
Ixil		60,000
2. *Greater Quichéan branch 26c*		
a. Uspanteco	NW Guatemala	15,000
b. *Quiché complex 10c*	C Guatemala	
Quiché (Achi)		680,000
Sacapultec		3,000
Sipacapa		3,000
Cakchiquel		434,000
Tzutujil		50,000
c. *Pocom complex 10c*	EC Guatemala	
Pocomam		30,000
Pocomchi		75,000
d. Kekchi		374,000
17. Tarasco isolate	SW Michoacán	72,000
18. Xinca complex 10c	SE Guatemala	
Eastern Xinca	Yupiltepeque, Jutiapa	extinct
Northern Xinca	Jumaytepeque	50
Southern Xinca	Chiquimulilla	100
Western Xinca	Guazacapan	100
19. Lencan family 20c		
Lenca	SW Honduras	25
Chilanga	E El Salvador	a few
20. Paya complex 10c	N Honduras	300
21. Misumalpan (Misuluan) family 43c		
A. Mosquito (Miskito)	Nicaragua, Honduras	115,000
B. *Matagalpa complex 10c*		
Matagalpa	Nicaragua, Honduras	100
Cacaopera	El Salvador	?
C. *Sumo complex 11c*		
Sumo, Úlua, Tahuajca	Nicaragua	200

*Indicates centuries of separation. †Not spoken in Meso-America. ‡There is some doubt whether these groups should be given the status of complexes.
§Sonoran languages spoken in Meso-America. ‖Varieties of the same language spoken in different countries (and having different names).

is 74 percent). The following are terms and categories for degree of relatedness, correlated with glottochronological time depths, that will be used to describe the various Meso-American language groups. The figures given are minimal bounds.

term	centuries of separation	percentage of cognates
dialects	0–5	86–100
language complex	7–11	71–81
language group	13–17	60–68
branch (or family if there is no superordinate category)	19–26	45–56
language family	35–45	26–35
stock or phylum	55–65	14–19

In Table 62 every family (group) and isolate has a separate number from 1 to 21. Each of the 21 headings specifies the name of a grouping, with alternative names. Numbers in parentheses following language names indicate that there are several closely related languages all referred to by the same name. For each language grouping the various levels of relatedness are specified, including glottochronological figures (c = centuries), which are Swadesh's, except for Mixe-Zoque, Mayan, and Xincan, which are those of the U.S. linguist Terrence Kaufman. Family and stock names are formed in the following ways: (1) A typical language, usually the most widely spoken, is suffixed with *-an* (*e.g.,* Mixtecan). (2) Two typical names are chosen and compounded (*e.g.,* Mixe-Zoque). (3) Parts of two or more language names are joined, and *-an* is suffixed (*e.g.,* Oto-Manguean, Oto-Pamean, Mis-Uluan/Misumalpan).

Group names end in *-an* if the groups are further subgrouped but do not end in *-an* if they are immediately divided into discrete languages.

The map gives the approximate geographic distribution of the 21 language groupings and isolates of Meso-America. None of the extinct undocumented languages is indicated. Except for some outliers, separate languages within a grouping are not localized. An outlier is a language that has been carried into a foreign cultural and linguistic context by migration; for example, Mangue is a Chiapanec outlier in Misumalpan territory, Subtiaba is a Tlapanec outlier in Misumalpan territory, Pipil is a Nahua (Aztec) outlier in Quichéan, Xinca, Lencan, and Misumalpan territories. Outliers

In the following paragraphs the numbers in parentheses refer to groupings in Table 62.

Uto-Aztecan (1). The Uto-Aztecan family consists of some 27 languages that are universally recognized to fall into eight groups or branches—the Plateau group, Tubatulabal, the Southern California branch, Hopi, the Piman group, the Yaquian branch, the Coran group, and the Nahuan group. Tubatulabal and Hopi contain just one language each. The first four groups are commonly, but not universally, recognized as forming a Shoshonean division within the family. None of the Shoshonean languages is spoken in Meso-America, and no distribution or population data is cited for them in Table 62 (see above *North American Indian languages*). There are two common ways of grouping the remaining languages, depending on the position assigned the Nahuan group. Either Nahuan is considered as separate and the rest as forming a Sonoran division, thereby producing three divisions—Shoshonean, Sonoran, and Nahuan—or else Nahuan is included within Sonoran, thereby producing a Shoshonean versus Sonoran dichotomy, which is the arrangement used in this article. Several scholars believe that the "division" concept is faulty here and that Uto-Aztecan contains eight groups and branches that are not to be further grouped in any special way.

Only some Sonoran languages are spoken in Meso-America (indicated by signs [§] in Table 62). The extinct Tubar belongs to the Yaquian branch, but whether to the Tarahumara complex, the Cáhita complex, or neither, is not clear. The Nahuan group includes the extinct Pochutec, formerly spoken on the coast of Oaxaca, Mexico, and poorly documented; Pochutec is clearly very divergent from the rest of the group. The Aztec complex is considered by some to be a single language with several dialects. The three Aztec languages were spoken within the Aztec Empire as it was constituted in 1519. Pipil speakers, who also refer to their language as *nawat,* were not a part of the Aztec culture and probably represent a Toltec expansion from several centuries earlier.

Various scholars' work on Uto-Aztecan

In 1859, Johann Karl Buschmann, a German philologist, correctly identified all the then-known Uto-Aztecan languages as forming a family. In 1883 a French philologist, Hyacinthe de Charencey, divided Uto-Aztecan into Oregonian (=Shoshonean) and Mexican (=Sonoran), and, in 1891, in the United States, anthropologist Daniel Brinton recognized Shoshonean and divided the Sonoran division (of this article) into Nahuatlan (=Nahuan) and Sonoran (=the Sonoran of this article minus Nahuan). Brinton's division was followed by the United States biologist John Wesley Powell in his classification of North American languages.

Buschmann in 1859 and United States anthropological linguist Edward Sapir in 1915 contributed to the comparative study of Uto-Aztecan by assembling sizable numbers of cognate sets.

A number of now-acculturated and racially absorbed Indian ethnic groups of northern Mexico are believed by many to have spoken Uto-Aztecan languages, although only the language names are known, and not the languages themselves. These are: Suma, Jumano, Lagunero, Cazcán, Tecuexe, Guachichil, and Zacatec.

Uto-Aztecan is generally accepted by specialists as related to the Kiowa-Tanoan family of North America and with it to form the Aztec-Tanoan stock (or phylum).

Cuitlatec (2). The now extinct Cuitlatec language has not been linked convincingly with any other language or family, though the idea that it might be related to Uto-Aztecan has been entertained.

The Hokan hypothesis (3–5). In 1919 two United States anthropologists, Roland Dixon and Alfred Kroeber, tried to improve on an older North American classification by reducing the multiplicity of language groupings in California (about 50) to a manageable number of families and stocks. Working over a period of several years, they developed the hypothesis that most California languages belong to one of two great groupings (called phyla or superstocks), Hokan and Penutian. The formulation was accepted and extended by others. Hokan included Shasta, Achumawi, Atsugewi, Chimariko, Karok, Yanan, Pomoan, Washo, Esselen, Yuman, Salinan, and Chumashan. By 1891/92 it had been suggested that Yuman, Seri (3), and Tequistlatec (4) were related. In 1915 the matter was re-examined in the light of the Hokan hypothesis, and it was concluded that all of the languages named above are related. Since then most scholars familiar with Yuman languages have believed that Seri and Yuman are related, and many who accept the Hokan hypothesis believe that Seri and Yuman form a special group within Hokan.

Jicaque (5), which is very poorly documented, though still spoken, has plain, aspirated, and glottalized stops (different varieties of consonant sounds), as do many Hokan languages. In 1953 it was suggested that Jicaque is a Hokan language. The general acceptance of the proposition may have been uncritical, because the available data on Jicaque is hardly reliable.

Documented dead languages

Extinct languages of northeast Mexico. All of the several languages once spoken in northeast Mexico and South Texas have become extinct. Documented languages of Mexico are: Coahuilteco, Comecrudo, Cotoname, Naolan, and Tamaulipec (or Maratino). Those of Texas are Karankawa (and Klamkosh), Atakapa, and Tonkawa. John Wesley Powell classified the first three as forming a Coahuiltecan family. The other Mexican languages were unknown until recently. Each of the three Texan languages was considered by Powell to be an isolate. In 1920 Coahuiltecan was redefined to include Karankawa and Tonkawa and to be coordinate with Hokan in a Hokan-Coahuiltecan (=Hokaltecan) superphylum.

Tlapanec (6). The Tlapanec complex was first correctly identified by Walter Lehmann, a German physician, in 1920. In 1925 Edward Sapir tried to establish Subtiaba as a Hokan language, proposing some Proto-Hokan reconstructions that could account for the Subtiaba forms. This classification is generally accepted. More recently, however, Calvin Rensch, a U.S. missionary linguist, tried to validate the Oto-Manguean hypothesis (see below) by means of full-scale phonological reconstruction. He believed Tlapanec to be Oto-Manguean; others considered it to be intermediate between Oto-Manguean and Hokan. It must be kept in mind that most of the specialists who have immersed themselves in the study of large numbers of American Indian languages believe that almost all of them are genetically related to one another. This relationship derives from a period, perhaps 20,000 to 30,000 years ago, when some of the languages were still spoken in Asia. With such a point of view, correct grouping (or degree of relationship) is a more interesting question than genetic relatedness.

Oto-Pamean (7). The Oto-Pamean stock contains four groups and complexes, Chichimec, Pamean, Matlatzinca, and Otomían, of which only the last two are spoken within Meso-America. The exact number of languages within the Otomí complex is not yet determined, though there seem to be four. Oto-Pamean was first correctly identified in 1892.

Popolocan (8). The Popolocan family (which might more appropriately be called Mazatecan) was correctly identified in 1926. The exact number of languages within the Mazatec complex has not yet been determined, though there are at least two.

Problem of classifying Amuzgo

Mixtecan (9). There is some difference of opinion as to how the various languages here included within Mixtecan are to be grouped. The main problem is whether Amuzgo is Mixtecan or a separate branch within Oto-Manguean. It has been included within Mixtecan in some systems and excluded from it in others. There seem to be three languages within the Mixtec group, a subdivision of Mixtecan.

Zapotecan (10). The Zapotecan family was correctly identified by William Mechling in 1912, but only Francisco Belmar, a Mexican philologist, correctly recognized that Papabuco is a separate language, neither Zapotec nor Chatino (in 1905). Belmar, however, incorrectly included Chinantec within Zapotecan. The Chatino language has several dialects. Within the Zapotec complex there are at least four languages, and perhaps more.

Chinantecan (11). The Chinantecan group contains approximately four languages, the exact number as yet undetermined. The separateness of Chinantecan within Oto-Manguean was recognized in 1912.

Manguean (12). The Manguean group was correctly identified by Belmar in 1905. Its members, formerly spoken in Chiapas (Mexico), and in Nicaragua, Honduras, and Costa Rica, are now extinct.

The Oto-Manguean hypothesis (7–12 or 6–13). Ever since 1891, it has been proposed that two or more of the above families (7–12) should be linked. Since about 1925,

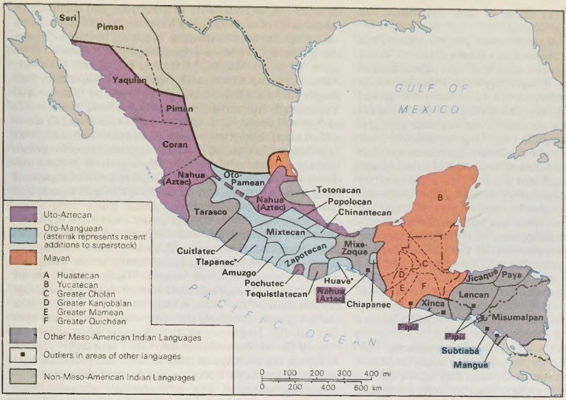


Figure 35: Distribution of Meso-American Indian languages c. AD 1500. Boundaries are schematic.

it has been generally accepted by specialists that the Oto-Pamean, Popolocan, Mixtecan, Zapotecan, Chinantecan, and Manguean groups form a larger genetic grouping (phylum), commonly labelled Oto-Manguean. This may be called the "classical Oto-Manguean formulation." Since 1950, work has been going on in the reconstruction of parent languages for each of the constituent families and groups. Since 1961, two revisions have been proposed in the formulation of what constitutes Oto-Manguean: the Tlapanec language complex has been recognized as included in or closely related to Oto-Manguean, and Huave has been proposed as an Oto-Manguean language. In the early 1970s, therefore, most Oto-Manguean specialists considered the grouping to consist of groups 6–13.

Reconstruction of Oto-Manguean phonology

The comparative study of the Oto-Manguean phylum has resulted in the first case in the Western Hemisphere in which the remote common ancestor of several language families has been phonologically reconstructed. Comparative linguistics at the phylum level has been largely unsuccessful with other postulated superstocks because of the relatively small number of cognates that can be identified. Except for Manguean, all Oto-Manguean languages are spoken in central Mexico.

Huave (13). Early proposals linked Huave to Mixe-Zoque and Mayan. Although this has not been generally accepted by many specialists, it has been uncritically repeated in most compilations. Recently, Morris Swadesh presented a reasonably well documented proposal for Huave as an Oto-Manguean language.

Mixe-Zoque (14). The Mixe-Zoque family consists of eight languages, which, comparative phonology and grammar suggest, form two branches—a Zoquean group, and a Mixean group including Tapachultec. Glottochronological figures, however, suggest a three-way division, as shown in the Table. The Mixe-Zoque family was correctly identified by Hyacinthe de Charencey in 1883. The Texistepec, Sayula, and Oluta languages of this family are all locally called Populuca.

Totonacan (15). The Totonacan family contains just two languages, of which one (Totonac) has at least three dialects. Possibly, Totonac is a complex.

Mayan (16). The Mayan family was correctly identified by a German ethnographer, Otto Stoll, in 1884. This family, with 24 languages and nearly 3,500,000 speakers, is the most diversified and populous language family of Meso-America. The Huastec language is separated by more than 1,000 miles from the nearest other Mayan language. Taken with the fact that the Huastecs did not share in the Classic Maya civilization, this requires a historical explanation involving the separation of Huastec from the rest of the family more than 2,500 years ago. Though the geographical extent of the Mayan languages is considerable, the Mayan peoples, languages, and cultures (as contrasted with those of the Aztecs), have never been particularly expansionist.

A number of attempts have been made to classify the Mayan languages, each one availing itself of more data than the last. The classification given here as of 1971 recognizes, at the lowest level, ten groupings. Specialists have disagreed on the precise positions of Tojolabal and Chuj, Motozintlec, Aguacatec, Uspantec, and Kekchí and have held no firm opinions about the Yucatec or Huastec complexes. Not much comparative work on the Mayan family has seen print, but much data has recently been collected. The main contributors to Mayan comparative studies have been the U.S. linguists Norman McQuown (1950s and 1960s) and Terrence Kaufman (1960s).

Hypotheses concerning superstocks

The Macro-Mayan and Macro-Penutian hypotheses. In 1931 L.S. Freeland, a U.S. anthropological linguist, tried to show that Mixe (Zoque) is related to the "Penutian" languages, a superstock that up until then had been limited to California, Oregon, Washington, and British Columbia. In 1935 it was suggested that the similarities between Uto-Aztecan, Tanoan, Kiowa, Penutian, Mixe-Zoque, and Mayan were such as to indicate the existence of a superstock, which it was proposed to call Macro-Penutian. This hypothesis had favour for a period but was never demonstrated nor taken very seriously by specialists. Since then the first three have been generally joined in Aztec-Tanoan. In 1942 it was suggested that Mixe-Zoque and Totonacan might be related genetically to each other and the two in turn might be related to Mayan, the resultant superstock to be called Macro-Mayan. Recently it has been claimed that Tarasco (17) probably belongs in Macro-Mayan as well, though the attempt to prove this has not been convincing to most Mayanists, to whom, minus Tarasco, the Macro-Mayan hypothesis seems as reasonable as the Hokan hypothesis.

Tarasco (17). Tarasco has been linked genetically by some not only to Marco-Mayan but also to both Zuni (in North America) and Quechua (in South America), but without general scholarly acceptance.

Xinca and Lencan (18–19). It has been suggested that Xinca and Lencan are related and that one or both of them is related to Mayan (16), Chibchan (in South America), or Uto-Aztecan (1). None of these hypotheses has been demonstrated as probable.

Languages outside Meso-America proper. The Paya language (20) and the Misumalpan family (21) are Central American languages spoken outside of the cultural area of Meso-America proper, though they have Meso-American outliers in their territory. Paya (20) has been linked in hypotheses to Chibchan and Cariban (both in South America), and perhaps to others, but not convincingly. The Misumalpan family (21) has been recognized since 1895. Since that date some scholars have believed that the three languages and complexes listed are coordinate, and others have believed that the first two constitute one group and the other consitutes a second group. Although the family relationship can be verified on inspection, no supporting comparative work has been published. Previous comprehensive classifications of the Meso-American Indian languages were presented by the U.S. anthropologists Cyrus Thomas and John R. Swanton in 1911 in *Indian Languages of Mexico and Central America and Their Geographical Distribution,* by Edward Sapir in the 14th edition of *Encyclopædia Britannica* (1929), and by Morris Swadesh in 1967 in *Handbook of Middle American Indians.*

Newly discovered languages and reconstructions. Although there are probably no uncharted areas in Meso-America, it is not necessarily the case that all the Indian languages of Meso-America have been correctly identified, and there are probably some multilingual Indian communities as well that are not known to be such. In 1967 Terrence Kaufman discovered a hitherto undocumented Mayan language spoken by several hundred Indians in four or five towns in southeast Chiapas and west central Guatemala. Although it appears to be closely related to Mam, Kaufman considered it a separate language and christened it Teco. Kaufman identified two more new Mayan languages in the course of a linguistic survey of Guatemala. These two new languages—Sacapultec (formerly considered Quiché) and Sipacapa (formerly assumed to be Mam)—are not documented in print and both belong to the Quiché complex.

Reconstruction of earlier forms of the Meso-American Indian languages has focussed primarily on phonology and vocabulary. Phonological and lexical comparative studies as well as reconstruction have been done for the following groups: Uto-Aztecan; Oto-Manguean—Oto-Pamean, Popolocan, Mixtecan, Zapotecan, Chinantecan, Manguean; Mixe-Zoque; and Mayan (in part). A small amount of grammatical comparison has been done within Oto-Manguean and Mixe-Zoque. In addition, some studies have been done of reconstructed vocabulary for the purpose of hypothesizing about the culture of the speakers of the protolanguages.

RELATION OF LANGUAGES TO HISTORICAL AND CULTURAL INFLUENCES

Pre-Columbian diffusion. The following are some of the important civilizations that have flourished in Meso-America:

civilization	period	location
Olmec	1200 BC–400 BC	Gulf Coast, Mexico
Monte Albán	400 BC–AD 700	Oaxaca, Mexico
Teotihuacán	AD 100–600	Central Mexico
Classic Maya	AD 300–900	Chiapas, Mexico; Petén, Guatemala
Toltec	900–1200	Central Mexico
Aztec	1300–1500	Central Mexico

The Aztecs spoke Nahuatl, as did the Toltecs. The Classic Maya probably spoke two or three Mayan languages, and the people of Monte Albán probably spoke one or more Zapotecan languages. No one knows what either the Teotihuacán people or the Olmecs spoke, but it has been surmised that at least some Olmecs spoke Mixe-Zoque languages and that the Teotihuacán people may have spoken Otomían languages (though an Aztec tradition says Totonac).

In the pre-Columbian period, there was naturally contact among Meso-American languages and occasional borrowing of vocabulary and other linguistic features. Partly because of the unavailability of grammars and dictionaries, actual cases of such diffusion have not been much studied.

Known contacts between various Indian groups

Some of the known contacts resulting in borrowing are the following: (1) Mixe-Zoque languages (Olmecs?) have given words to Mayan, Mixtecan, Zapotecan, Otomían, Aztec, Lencan, Xinca, and Jicaque; (2) Zapotecan languages (Monte Albán) have given words to Huastec and Yucatec; (3) Mayan languages (Mayas) have given words to Xinca, Lencan, and Jicaque; and (4) Nahuatl (Toltecs and Aztecs) has given words to Mayan, Lencan, other Uto-Aztecan languages, as well as to other Meso-American languages. Words diffused from these sources provide evidence that contact took place. Scholars know that contact must have taken place at particular times and places, and therefore can form hypotheses about where certain languages may have been spoken in the more remote past.

External relationships and contacts. Various scholars have suggested that some Meso-American language or family is related to a language or family (other than Uto-Aztecan) outside of Meso-America. These suggestions are mostly parts of larger attempts to synthesize the language classification of the New World, or of the whole world, and are usually based on the sometimes unexpressed view that all the languages of the Western Hemisphere or even of the whole world are ultimately genetically related. Although the assumption may be true, the proposed connections have been unconvincing to specialists in Meso-American languages. The only generally accepted larger groupings are Hokan and Penutian. Most scholars do not have the breadth of knowledge to be able to evaluate these vast proposals.

One proposal of external relationship probably has some merit. In 1961 it was suggested that Chipaya—a language spoken on the shores of Lake Titicaca in Bolivia—is genetically related to the Mayan languages. The hypothesis, proposed by Ronald Olson, a U.S. missionary linguist, was based on 120 sets of lexical comparisons between Chipaya and Proto-Mayan. The data cited are subject to more than one interpretation, because many of the comparisons involve semantic notions and word forms that are widespread in the Western Hemisphere; also, Chipaya has been so influenced grammatically by Aymara (which all Chipayas can speak) that any grammatical peculiarities it may once have shared with Mayan have disappeared. Because a core of data showing regular sound correspondences remains, it is probably necessary to assume that there is a historical connection between Chipaya and Mayan, possibly, but not demonstrably, a genetic relationship. The connection may have been direct—presumably from Meso-America to Bolivia via land—or there may be other languages in western South America that show prehistoric contacts with Mayan. The acceptance of a prehistoric linguistic connection, neither extremely remote nor extremely recent, between Meso-America and the Andes is quite provocative, inasmuch as other evidence exists for early culture contact between Meso-America and the Andes, Meso-America generally being the donor and the Andes generally being the beneficiary; *e.g.,* in the case of corn. Later diffusion from South America to Meso-America also occurred; *e.g.,* witness the transference of peanuts, metallurgy, hammocks.

Pre-historic linguistic connection between Meso-America and the Andes

Interaction between Spanish and Indian languages. In modern Meso-America, the dominant European language is Spanish. The speakers of all Meso-American Indian languages include some who are bilingual; and a few languages are spoken by almost totally bilingual populations. Most Indian languages spoken by sizable populations have at least 50 percent monolingual speakers. All Meso-American languages with a significant number of bilingual speakers have been influenced by Spanish, primarily in the areas of vocabulary, particles, and word order. Since the Spanish conquest, Meso-American languages have been borrowing words from Spanish, and, because the kind of Spanish spoken has changed somewhat over the years, both in vocabulary and pronunciation, different historical periods are usually distinguishable in lexical borrowings. For a variety of reasons, certain function words, primarily conjunctions and adverbs, are frequently borrowed from Spanish; *e.g., ya* "already," *pero* "but," *hasta* "until," *y* "and," *o* "or," *ni* "not even," *hasta* "even," *si* "if," *cuando* "when," *porque* "because," *por eso* "therefore, so," *entonces* "then." Some languages have assimilated the Spanish word order of subject–verb–object.

Conversely, the Spanish of Meso-America has been the recipient of vast amounts of lexical material from local languages, primarily Nahuatl. The borrowing has provided names of plants, animals, artifacts, and social forms indigenous to Meso-America and lacking names in Spanish. Among the reasons that Nahuatl has been the primary source is that the Aztecs were the first Meso-American people conquered by the Spaniards; the Aztecs had outposts in many parts of Meso-America; the Spaniards recruited Aztecs, particularly as guides, into their military force to assist their venture of subduing the rest of Meso-America; and, for several decades, Aztec, written in Roman orthography, was used in many parts of Meso-America to keep official records, such as deeds, wills, and censuses.

Many of the words borrowed into Spanish from Aztec have since passed in turn into English; *e.g.*, chili, chile, or chilli (Spanish *chile*), avocado (Spanish *aguacate*), chicle, chocolate, peyote, coyote, tomato (Spanish *tomate*), ocelot (Spanish *ocelote*), guacamole, mescal.

Bilingualism among Indians

In some parts of Meso-America, because of economic and social conditions, an Indian may speak one or more Indian languages besides his own. This is common in Guatemala, where some areas have been recently colonized by speakers of more than one language, or some communities have received outside settlers in the more remote past.

The names used in this article for the Meso-American Indian languages are English versions of the Spanish terms for them. Only in a few cases are these names the ones actually used by the people who speak the languages in question. First, most of the names are of Aztec origin, because at the outset the Spanish learned of local phenomena primarily via Aztec. Secondly, some languages have no special name of their own, simply being called "our language."

Pre-Columbian writing. Most of the Meso-American cultures shared a mathematical notation and calendrical system that had been developed and diffused in the distant past, probably before 500 BC. At the time of European contact the Aztecs, Zapotecs, Mixtecs, Otomís, Mayans, and perhaps some others were all producing records on stone (inscriptions) and on a type of homegrown paper (produced from the amate tree, *Ficus glabrata*), these latter being commonly called codices. Except for the Mayan system, which probably originated before AD 1, the records cannot properly be called writing, in that it was not possible to represent all of speech, but only numbers, dates, and names (pictographically). The Mayan system, besides representing all these, was also used to represent morphemes (words and word elements) and phonemes (distinctive sounds). Presumably the symbols used in this system (called glyphs) represent individual phonemes, syllables, and morphemes; and they give semantic information as well to take the ambiguity out of homophonous readings. Several scholars have devoted much time to the study of Mayan writing, but, to date, the results have not been very impressive. A few scholars outside the Meso-American field believe the Mayan writing system is purely ideographic and hence inherently undecipherable without a bilingual inscription or text in a known language. All specialists within the Mayan field hold that the Mayan is a mixed ideographic and phonological system.

What may be delaying progress in the deciphering of Mayan writing is the absence of reconstructions for intermediate groupings within the Mayan family (*e.g.*, Proto-Yucatecan, Proto-Cholan, and others) and ignorance of Mayan languages other than colonial Yucatec on the part of the investigators. Efforts are being made to correct these deficiencies, particularly by Mexican specialists. It is not known whether Mayan writing was used to write more than one language and, if so, what the languages were. If only one, it was probably either Proto-Cholan or Proto-Yucatecan. The symbols used in all the pre-Columbian notation systems are obviously pictographic in origin, as was the case in the ancient Egyptian, Sumerian, ancient Chinese, and Indus Valley writing systems.

LINGUISTIC CHARACTERISTICS

In general, all the languages of a particular family are typologically similar to one another both in phonology and grammar. Among the 21 language groupings in Meso-America, there are several types of sound systems and grammatical systems. Because study in this area has hardly begun, nothing very secure can be asserted here, but some general characteristics can be outlined on the basis of data for the following reasonably well-documented languages: Tequistlatec, Otomí, Mazatec, Mixtec, Zapotec, Chinantec, Aztec, Zoque, Totonac, Quiché, and Tarasco.

Diverse types of sound systems and grammars

Phonologically, there is a wide diversity among Meso-American languages. Voiced spirants—*i.e.*, sounds like English *v*, *z*, or *th* in "then"—are missing from all Meso-American languages. Other phonological features in these languages include a voiceless lateral spirant sound, *lh* (in Tequistlatec and Totonac); a lateral affricate, *tl* (in Aztec and Totonac); a postvelar stop, *q*, in contrast with a velar stop, *k* (in Quiché and Totonac); glottalized vowels (in Zapotec, Zoque, Aztec, and Totonac); glottalized consonants (in Tequistlatec, Quiché, Otomí, and Mazatec); aspirated stops (in Tarasco, Otomí, and Mazatec); voiced stops (in Tequistlatec, Otomí, Mazatec, and Chinantec); prenasalized stops (in Otomí, Mazatec, and Mixtec); nasalized vowels (in Otomí, Mazatec, Mixtec, and Chinantec); a labiovelar stop, *kw*, sometimes contrasting with a bilabial stop, *p* (in Otomí, Mazatec, Mixtec, Aztec); tone and stress accent (tone in Otomí, Mazatec, Mixtec, Chinantec, Zapotec; stress in Tarasco and Tequistlatec); and initial and final consonant clusters (in Tequistlatec).

Grammatically, Meso-American languages are rather diverse, but, according to available data, they fall into three main types: Type A, an Oto-Manguean type, is rightward expanding (*i.e.*, modifiers follow the elements they modify) and synthetic to a low degree (*i.e.*, characterized by relatively few morphemes per word). It employs prefixes and prepositions, and it seldom uses compounding to form words. Type B, an intermediate type, is prepositional, like A, and averagely synthetic, making some use of prefixes (subjects, objects, and possessors) and much use of suffixes. It is mildly leftward expanding (*i.e.*, modifiers precede the elements they modify) and is mainly represented by Mayan and Uto-Aztecan languages but partially by Mixe-Zoque and Totonacan. Type C, a leftward expanding type, is highly synthetic with great use of suffixes and postpositions and active ablaut (an interchange among consonants and vowels for the purpose of derivation or inflection). It is represented by Tarasco and, partially, by Totonacan and Mixe-Zoque.

Grammatical generalizations

There are a number of grammatical generalizations that can be made about all, or most, Meso-American Indian languages. (1) The genitive relationship between nouns or noun phrases is (except for Tarasco) expressed by means of a possessive pronoun with the possessed noun; *e.g.*, "the dog's fleas" is expressed as "his fleas the dog." (2) Locative notions, such as "above," "below," "in," "on," "beside," are not expressed by prepositions and adverbs, as in European languages, but by means of location nouns (meaning "aboveness," "belowness," "belly," "surface," "side," and so forth), which are always combined with a possessive pronoun, the function of which is to indicate the "object" of the prepositional–adverbial notion. Most languages, however, have at least one generic relational particle that is combined in a phrase with a location noun and its object and has "generic prepositional" function; thus "on the table" is expressed "at (generic particle) its-top the table," or "in the box" is expressed "at its-inside the box." Whereas in most languages the generic relational particles are prepositions, Zoque and Tarasco have postpositions, which are in part related to location nouns.

(3) Within the verbal system, aspect (type of action—*e.g.*, ongoing, habitual, finished, potential, and so forth) is well developed, and tense (time—*e.g.*, now, in the past, in the future) is generally weakly developed. (4) The copula, or equational verb "be," is not expressed in most Meso-American languages. (5) Case suffixes are generally absent,

being present in just three languages: Tarasco has a genitive case, an objective case, and various locational cases; Aztec and Zoque have only locational cases, and these are usually related to location nouns. (6) A relative clause that modifies a noun follows it in all the languages of the sample above; *e.g.*, "the man whom I saw (on the street yesterday)." (7) Some Oto-Manguean languages and some Mayan languages distinguish an inclusive pronoun "we" ("I and you") from an exclusive "we" ("I and he/they"). (8) Gender, or inflectional agreement of other word classes in the noun phrase with the noun itself, is rare in Meso-American languages and is limited to some Oto-Manguean languages. (9) Noun subclassification in the context of possession is not uncommon. In some languages, some nouns undergo form changes when possessed; these languages, therefore, have at least two classes of nouns. In other languages, the possessive pronouns differ in form according to how they are associated with different classes of nouns. In languages in which the semantic motivation for such a subdivision is clear, the main kind of distinction is between intimate possession (body parts, kinship terms, articles of clothing) and casual possession (domestic animals, tools). (10) Some languages (Mayan, Mixe-Zoque) distinguish between the subject (actor) of a transitive verb and that of an intransitive verb by the form of the associated affixed pronoun. (11) Most Meso-American languages average more than one morpheme per word, and Tarasco and Totonac average more than two morphemes per word. (12) Most Meso-American languages (except Aztec) have consonantal or vocalic ablaut, or else show in their vocabulary sets of words that seem to be related through a formerly functional ablaut system. (13) The numeral systems are vigesimal–decimal; that is, counting is from 1 to 10, then from 11 to 20, then from 21 to 40 (adding 1–20 to 20), then from 41–60 (adding 1–20 to 40), and so on, with special terms for 400 (20×20), 8,000 ($20 \times 20 \times 20$), 160,000 ($20 \times 20 \times 20 \times 20$), and so on. In most languages (except Mayan) the numeral expressions for 6 through 9 (sometimes 5 through 9) are compounds of $5 + 1$, $5 + 2$, $5 + 3$, $5 + 4$, or the like. (14) In all the languages referred to here, a numeral precedes the noun it quantifies. (Te.K.)

South American Indian languages

South American Indian languages once covered and today still partially cover all of South America, the Antilles, and Central America to the south of a line from the Gulf of Honduras to the Nicoya Peninsula in Costa Rica. Estimates of the number of speakers in that area in pre-Columbian times vary from 10,000,000 to 20,000,000. In the early 1980s there were approximately 15,900,000, more than three-fourths of them in the central Andean areas. Language lists include around 1,500 languages, and figures over 2,000 have been suggested. For the most part, the larger estimate refers to tribal units whose linguistic differentiation cannot be determined. Because of extinct tribes with unrecorded languages, the number of languages formerly spoken is impossible to assess. Only between 550 and 600 languages (about 120 now extinct) are attested by linguistic materials. Fragmentary knowledge hinders the distinction between language and dialect and thus renders the number of languages indeterminate.

Number of languages and speakers

Because the South American Indians originally came from North America, the problem of their linguistic origin involves tracing genetic affiliations with North American groups. To date only Uru-Chipaya, a language in Bolivia, is surely relatable to a Macro-Mayan phylum of North and Meso-America. Hypotheses about the probable centre of dispersion of language groups within South America have been advanced for stocks like Arawakan and Tupian, based on the principle (considered questionable by some) that the area in which there is the greatest variety of dialects and languages was probably the centre from which the language groups dispersed at one time; but the regions in question seem to be refugee regions, to which certain speakers fled, rather than dispersion centres.

South America is one of the most linguistically differentiated areas of the world. Various scholars hold the plausible view that all American Indian languages are ultimately related. The great diversification in South America, in comparison with the situation of North America, can be attributed to the greater period of time that has elapsed since the South American groups lost contact among themselves. The narrow bridge that allows access to South America (*i.e.*, the Isthmus of Panama) acted as a filter so that many intermediate links disappeared and many groups entered the southern part of the continent already linguistically differentiated.

Investigation and scholarship. The first grammar of a South American Indian language (Quechua) appeared in 1560. Missionaries displayed intense activity in writing grammars, dictionaries, and catechisms during the 17th century and the first half of the 18th. Data were also provided by chronicles and official reports. Information for this period was summarized in Lorenzo Hervás y Panduro's *Idea dell' universo* (1778–87) and in Johann Christoph Adelung and Johann Severin Vater's *Mithridates* (1806–17). Subsequently, most firsthand information was gathered by ethnographers in the first quarter of the 20th century. In spite of the magnitude and fundamental character of the numerous contributions of this period, their technical quality was below the level of work in other parts of the world. Since 1940 there has been a marked increase in the recording and historical study of languages, carried out chiefly by missionaries with linguistic training, but there are still many gaps in knowledge at the basic descriptive level, and few languages have been thoroughly described. Thus, classificatory as well as historical, areal, and typological research has been hindered. Descriptive study is made difficult by a shortage of linguists, the rapid extinction of languages, and the remote location of those tongues needing urgent study. Interest in these languages is justified in that their study yields basic cultural information on the area, in addition to linguistic data, and aids in obtaining historical and prehistorical knowledge. The South American Indian languages are also worth studying as a means of integrating the groups that speak them into national life.

Research on the languages

Classification of the South American Indian languages. Although classifications based on geographical criteria or on common cultural areas or types have been made, these are not really linguistic methods. There is usually a congruence between a language, territorial continuity, and culture, but this correlation becomes more and more random at the level of the linguistic family and beyond. Certain language families are broadly coincident with large culture areas—*e.g.*, Cariban and Tupian with the tropical forest area—but the correlation becomes imperfect with more precise cultural divisions—*e.g.*, there are Tupian languages like Guayakí and Sirionó whose speakers belong to a very different culture type. Conversely, a single culture area like the eastern flank of the Andes (the Montaña region) includes several unrelated language families. There is also a correlation between isolated languages, or small families, and marginal regions, but Quechumaran (Kechumaran), for instance, not a big family by its internal composition, occupies the most prominent place culturally.

Most of the classification in South America has been based on inspection of vocabularies and on structural similarities. Although the determination of genetic relationship depends basically on coincidences that cannot be accounted for by chance or borrowing, no clear criteria have been applied in most cases. As for subgroupings within each genetic group, determined by dialect study, the comparative method, or glottochronology (also called lexicostatistics, a method for estimating the approximate date when two or more languages separated from a common parent language, using statistics to compare similarities and differences in vocabulary), very little work has been done. Consequently, the difference between a dialect and language on the one hand, and a family (composed of languages) and stock (composed of families or of very differentiated languages) on the other, can be determined only approximately at present. Even genetic groupings recognized long ago (Arawakan or Macro-Chibchan) are probably more differentiated internally than others that have been questioned or that have passed undetected.

Extinct languages present special problems because of poor, unverifiable recording, often requiring philological interpretation. For some there is no linguistic material whatsoever; if references to them seem reliable and unequivocal, an investigator can only hope to establish their identity as distinct languages, unintelligible to neighbouring groups. The label "unclassified," sometimes applied to these languages, is misleading: they are unclassifiable languages.

Problem of language names

Great anarchy reigns in the names of languages and language families; in part, this reflects different orthographic conventions of European languages, but it also results from the lack of standardized nomenclature. Different authors choose different component languages to name a given family or make a different choice in the various names designating the same language or dialect. This multiplicity originates in designations bestowed by Europeans because of certain characteristics of the group (*e.g.*, Coroado, Portuguese "tonsured" or "crowned"), in names given to a group by other Indian groups (*e.g.*, Puelche, "people from the east," given by Araucanians to various groups in Argentina), and in self-designations of groups (*e.g.*, Carib, which, as usual, means "people" and is not the name of the language). Particularly confusing are generic Indian terms like Tapuya, a Tupí word meaning enemy, or Chuncho, an Andean designation for many groups on the eastern slopes; terms like these explain why different languages have the same name. In general (but not always), language names ending in *-an* indicate a family or grouping larger than an individual language; *e.g.*, Guahiboan (Guahiban) is a family that includes the Guahibo language, and Tupian subsumes Tupí-Guaraní.

There have been many linguistic classifications for this area. The first general and well-grounded one was that by U.S. anthropologist Daniel Brinton (1891), based on grammatical criteria and a restricted word list, in which about 73 families are recognized. In 1913 Alexander Chamberlain, an anthropologist, published a new classification in the United States, which remained standard for several years, with no discussion as to its basis. The classification (1924) of the French anthropologist and ethnologist Paul Rivet, which was supported by his numerous previous detailed studies and contained a wealth of information, superseded all previous classifications. It included 77 families and was based on similarity of vocabulary items. Čestmír Loukotka, a Czech language specialist, contributed two classifications (1935, 1944) on the same lines as Rivet but with an increased number of families (94 and 114, respectively), the larger number resulting from newly discovered languages and from Loukotka's splitting of several of Rivet's families. Loukotka used a diagnostic list of 45 words and distinguished "mixed" languages (those having one-fifth of the items from another family) and "pure" languages (those that might have "intrusions" or "traces" from another family but totalling fewer than one-fifth of the items, if any). Rivet and Loukotka contributed jointly another classification (1952) listing 108 language families that was based chiefly upon Loukotka's 1944 classification. Important work on a regional scale has also been done, and critical and summarizing surveys have appeared.

Current classifications

Current classifications are by Loukotka (1968); a U.S. linguist, Joseph Greenberg (1956); and another U.S. linguist, Morris Swadesh (1964). That of Loukotka, based fundamentally on the same principles as his previous classifications, and recognizing 117 families, is, in spite of its unsophisticated method, fundamental for the information it contains. Those of Greenberg and Swadesh, both based upon restricted comparison of vocabulary items but according to much more refined criteria, agree in considering all languages ultimately related and in having four major groups, but they differ greatly in major and minor groupings. Greenberg used short lexical lists, and no evidence has been published in support of his classification. He divided the four major groups into 13 and these, in turn, into 21 subgroups. Swadesh based his classification upon lists of 100 basic vocabulary items and made groupings according to his glottochronological theory (see above). His four groups (interrelated among themselves and with groups in North America) are subdivided into 62 subgroups, thus, in fact, coming closer to more conservative classifications. The major groups of these two classifications are not comparable to those recognized for North America, because they are on a more remote level of relationship. In most cases the lowest components are stocks or even more distantly related groups. It is certain that far more embracing groups than those accepted by Loukotka can be recognized—and in some cases this has already been done—and that Greenberg's and Swadesh's classifications point to many likely relationships; but they seem to share a basic defect, namely, that the degree of relationship within each group is very disparate, not providing a true taxonomy and not giving in each case the most closely related groups. On the other hand, their approach is more appropriate to the situation in South America than a method that would restrict relationships to a level that can be handled by the comparative method.

At present, a true classification of South American languages is not feasible, even at the family level, because, as noted above, neither the levels of dialect and language nor of family and stock have been surely determined. Beyond that level, it can only be indicated that a definite or possible relationship exists. In the accompanying chart—beyond the language level—recognized groups are therefore at various and undetermined levels of relationship. Possible further relationships are cross-referenced. Of the 82 groups included, almost half are isolated languages, 25 are extinct, and at least 10 more are on the verge of extinction. The most important groups are Macro-Chibchan, Arawakan, Cariban, Tupian, Macro-Ge, Quechumaran, Tucanoan, and Macro-Pano-Tacanan.

Macro-Chibchan. Macro-Chibchan languages, which form the linguistic bridge between South and Central America, are spoken from Nicaragua to Ecuador. Spread compactly in Central America and in western Colombia and Ecuador, they include approximately 40 languages spoken by more than 400,000 speakers. The group is probably more differentiated than a stock, languages not belonging to Chibchan being strongly differentiated. In the Colombian Andes a now extinct Chibchan language was the language of the highly developed Muisca culture. Important present-day languages include Guaymí (about 20,000 speakers) and Move (about 15,000) in Panama, Cuna (600) and Páez (37,000) in Colombia, and Cayapa, or Colorado (4,000), in Ecuador. A connection with Cariban has been suggested, and it is possible that such a relationship could be found through Warao (Warrau) and Waican (Waikan) on the one hand and through Chocó (Cariban) on the other.

Extent of the Arawakan languages

Arawakan. Arawakan languages formerly extended from the peninsula of Florida in North America to the present-day Paraguay–Argentina border, and from the foothills of the Andes eastward to the Atlantic Ocean. More than 55 languages are attested, many still spoken. Around 40 groups still speak Arawakan languages in Brazil, and others are found in Peru, Colombia, Venezuela, Guyana, French Guiana, and Surinam. Taino predominated in the Antilles and was the first language to be encountered by Europeans; although it rapidly became extinct, it left many borrowings. As did most languages of the tropical forest, the Arawakan languages receded with the influx of Spanish and Portuguese, mainly through group extinction; thus, 14 groups became extinct in Brazil between 1900 and 1957. Important languages still spoken are Goajiro (52,000 speakers) in Colombia, Campa (41,000) and Machiguenga (11,000) in Peru, and Mojo (more than 15,000) and Bauré (4,500) in Bolivia. Although most Arawakan languages have been recognized as such for a long time, they are greatly differentiated. They are most probably related to both the Macro-Pano-Tacanan and Macro-Mayan language groups.

Cariban. Cariban languages, numbering approximately 50, were spoken chiefly north of the Amazon but had outposts as far as the Mato Grosso in Brazil. The group has undergone drastic decline, and only about 22,000 people speak Cariban languages today, mostly in Venezuela and Colombia; they have disappeared from the Antilles and have been much reduced in Brazil and the Guianas. The

Table 63: South American Indian Language Groups*

language	location
1. Alacalufan (47): Aksanas or Kaueskar, Alacaluf, Caucau or Caucawe	Chile
2. Andoque (9)	Colombia
3. Araucanian or Mapuche (32, 37)	Chile, Argentina
4. Arawakan (43, 44)	
A. Amuesha	Peru
B. Apolista or Lapachu†	Bolivia
C. Arauan: Araua†, Curina, Madihá, Paumarí, Yamadí	Brazil
D. Chamicuro	Peru
E. Maipurean	
1. Achagua, Amarizana, Capite Minanei, Cauyarí, Guarú, Guayupé, Maipure†, Piapoco, Resígaro, Tariana, Warakena†, Yucuna	Colombia
Anauya, Baré, Curipaco, Guinau, Mandawaca, Paraujano [Parauhano]	Venezuela
Araikú†, Aruan†, Cariay†, Carútana, Catapolítani, Cawishana†, Hohodene, Manao†, Mapanai, Marawá†, Mariate†, Maulieni, Moriwene, Pasé†, Siusí, Wainumá†, Wiriná, Yabaana, Yumana	Brazil
Arawak or Lokono	Guyana, Fr. Guiana
Goajiro	Colombia, Venezuela
Adzaneni, Ipeca	Brazil, Colombia
Island Carib	Dominica
2. Atorai, Mapidian	Guyana
Wapishana	Guyana, Brazil
3. Baníva, Yavitero	Venezuela
4. Bauré, Mojo or Ignaciano, Muchojeone, Pauna, Paicone	Bolivia
5. Campa, Machiguenga, Piro	Peru
Canamari, Chontaquiro [Chontakiro], Cuniba†, Cushineri†, Ipuriná	Brazil
Inapari	Bolivia
6. Caripuna, Marawan	Brazil
7. Chané†	Argentina
Guaná	Paraguay, Brazil
Quiniquiano, Tereno	Brazil
8. Paressí	Brazil
Sarave	Bolivia
F. Taino†	Antilles
G. Morique [Morike] or Mayoruna	Peru
5. Atacama or Cunza or Lincan Antai†	Argentina, Chile
6. Auake or Arutani	Brazil, Venezuela
7. Auishiri [Awishira]	Peru
8. Baenan	Brazil
9. Bora-Huitotoan (2)	
A. Boran: Bora or Miraña, Emejeite, Muinane	Colombia
B. Huitotoan [Witotoan]:	
Andoquero, Huitoto [Witoto]	Colombia
Coeruna†	Brazil
Ocaina, Orejone†, Nonuya or Achote [Achiote]	Peru
10. Canichana	Bolivia
11. Capixana or Canoe	Brazil
12. Cariban (70)	
A. 1. Acawai (Arecuna, Camaracoto, Ingarico) Waica	Venezuela
Taulipang or Ipuricoto or Pemon	Brazil, Venezuela
2. Apalai, Aracajú, Upurui	Brazil
Oyana	Suriname
Rucuyen	Fr. Guiana
3. Apiacá or Apingi†, Arara†, Parirí†	Brazil
4. Atroahi (Yauaperi), Quirixana, Waimiri [Waimiry]	Brazil
5. Bakairi† [Bacairi], Nahukua (Nahucua), Yaruma†	Brazil
6. Bonari†, Hishkariana, Parucoto, Waiboi, Waiwai	Brazil
7. Cachuene or Caxuiana, Mutuan†, Pauxí†, Saluma, Wayewé	Brazil
Chiquena [Chikene] or Shikiana	Brazil, Guyana
8. Carare, Opone	Colombia
9. Caribe or Calina or Galibi	Antilles, Guianas
Carif	Belize, Honduras
10. Carijona (Guaque [Guake], Umaua)	Colombia
11. Colima†, Muzo, Pijiao	Colombia
12. Cumanagoto† (Chayma†, Tamanaco), Tivericoto†	Venezuela
13. Keseruna, Macushí, Paraviyana†, Purukoto†, Zapara	Brazil
14. Mariquitare or Decuana	Brazil, Venezuela
Yecuana or Mayongong (Cunuana, Ihuruana)	Venezuela
15. Mapoyo or Nepoyo, Yauarana	Venezuela
16. Motilon	Colombia, Venezuela
17. Palmela† [Palmella]	Brazil
18. Panare	Venezuela
19. Patagon	Peru
20. Pawishana	Brazil
21. Pianocoto	Brazil
Tliometesen	Suriname
Trio (Ocomayana, Urucuyena [Urucuena], Wama)	Suriname, Brazil
22. Pimenteira†	Brazil
23. Yao†	Fr. Guiana, Trinidad
B. Chocó: Chamí, Sambú [Sambo], Waunana	Colombia
13. Cariri or Kiriri	Brazil
14. Catacao†: Catacao†, Colan	Peru
15. Catembri or Mirandela†	Brazil
16. Catuquina [Catukina]: Bendiapa (Canamari, Parawa), Catuquina [Catukina] or Wiridyapa, Catauxi, Tucundiapa [Tucundyapa]	Brazil
17. Cayuvava (70)	Bolivia
18. Chapacura: Chapacura or Huachi, Itene or Moré, Itoreauhip, Nape, Pacahanovo, Quitemo†	Bolivia
Cumaná (Abitana), Torá, Urupá (Yarú [Jarú]), Wañám [Wanyam]	Brazil
19. Chiquito or Tarapecosi (42)	Bolivia
20. Cholonan: Cholona or Seeptsá†, Híbito†	Peru
21. Cofán	Colombia, Ecuador
22. Culli or Ilinga†	Peru
23. Erikbaktsa or Canoeiro	Brazil
24. Gamela†	Brazil
25. Gorgotoqui	Brazil
26. Guahiboan (4): Chiricoa, Guahibo	Venezuela, Colombia
Churuya, Guayabero	Venezuela
27. Guaycurú-Charruan	
A. Charruan†: Chaná†	Uruguay
Charrúa† (Guenoa or Minuan†)	Uruguay, Argentina, Brazil
B. Guaycuruan: Abipón or Callaga†	Argentina, Paraguay
Caduveo or Mbayá or Guaycurú	Brazil, Argentina, Paraguay
Guachí†	Brazil
Mocoví	Argentina
Payaguá or Lengua	Paraguay
Toba-Pilagá	Argentina, Bolivia
28. Guamo†	Venezuela
29. Guató	Bolivia, Brazil
30. Guennaken or Gununa-Kune or Puelche†	Argentina
31. Huarian: Huari or Corumbiara, Masaca or Aicana	Brazil
32. Huarpean† (37): Allentiac†, Millcayac†	Argentina
33. Iranxe (4)	Brazil
34. Jirajaran: Ayomán, Gayón†, Jirajara	Venezuela
35. Kaliana or Sape	Venezuela
36. Koaia	Brazil
37. Jebero-Jivaroan	
A. Jeberoan or Cahuapanan: Cahuapana or Chuncho [Concho], Chayavita, Jebero [Chébero], Miquirá, Yamorai	Peru
B. Jívaroan: Aguaruna	Peru
Jívaro or Shuara (Achual, Huambisa [Wambisa])	Ecuador
38. Kukura†	Brazil
39. Leco	Bolivia
40. Lulean (48)	
A. Lule or Tonocoté†	
B. Vilela or Chunupí (Atalalá, Ocolé, Uacambabelé)†	Argentina
41. Macro-Chibchan	
A. Chibchan	
1. Abiseta or Orosi or Tucurrique [Tucurrike], Boruca or Turucaca†, Bribrí or Lari, Cabecar, Chiripó, Estrella†, Terraba or Brurán†, Tirub or Rayado†	Costa Rica
2. Andaquí	Colombia
3. Atanque or Busintana, Bintucua or Ijca, Cágaba or Koghi, Guamaca or Arsario, Tairona or Teyuna†	Colombia
4. Bairira or Cunaguasaya, Motilón or Dobocubí	Colombia
Mape	Venezuela
5. Betoi†	Colombia
6. Cara or Imbaya†, Cayapa or Nigua, Colorado or Campaz or Colima or Satxila, Pasto†	Ecuador
Cuaiquier, Muellama†, Telembi†	Colombia
7. Catio†, Nutabé†	Colombia
8. Chibcha or Muisca†, Tunebo	Colombia
9. Chimila, Malibú†	Colombia
10. Changuena†, Chumula† (Gualaca†)	Panama
11. Coconuco, Guambiana or Silviano, Moguex, Totoró	Colombia
12. Corobisi, Guatuso, Guetar, Suerre or Camachi	Costa Rica
13. Cuna	Panama
Cueva	Colombia
14. Guaymí, Move (Penonemeño†)	Panama
15. Panzaleo or Quito†	Ecuador
Paez	Colombia
16. Rama	Nicaragua
17. Sebondoy or Kamsá or Coche or Mocoa, Quillasinga†	Colombia
B. 1. Esmeralda or Atacame†	Ecuador
2. Yaruro	Venezuela
C. Itonama	Bolivia
D. Paya	Honduras
E. Sumo-Mosquito-Matagalpa	
1. Matagalpa† (Cacaopera)	Ecuador, Suriname
Jinotega or Chingo	Nicaragua
2. Mosquito [Miskito]	Honduras, Nicaragua
Ulua, Sumo	Nicaragua
F. Waican or Yanoaman: Karimé [Carimé], Pakidai-Surara, Paucosa	Brazil
Sanemá or Samatari (Pubmatari)	Venezuela
Shamatari, Shirianá or Casapare, Waica [Waika]	Brazil, Venezuela
G. Warao [Warrau] or Guarauno	Venezuela
42. Macro-Ge (70)	
A. Bororoan	
1. Bororo	Brazil
2. Otuké	Bolivia
B. Botocudo or Aymoré†	
C. Fulnió†	Brazil
D. Ge	
1. Akroá†, Xakriaba† [Shacriaba], Xavante [Shavante], Xerente [Sherenté]	Brazil
2. Apinayé-Kayapó, Eastern Timbira, Suyá	Brazil
3. Caingang [Kaingang], Xokleng [Shocleng]	Brazil
4. Southern Kayapó	
E. Jeikó [Jeico]†	Brazil
F. Kamakán†	Brazil
G. Karajá	Brazil
H. Kapoxo† (Kumanaxo†), Malalí†, Maxakalí†, Monoxo†, Patashó†	Brazil
I. Ofayé or Opayé-Shavante†	Brazil
J. Purí-Coroado†: Coroado†, Coropó†, Purí†	
43. Macro-Mayan (4): Uru-Chipaya (Uru, Chipaya)	Bolivia
44. Macro-Pano-Tacanan (4)	
A. Chon: Haush or Manekenken†, Ona or Shelknam†, Tehuelche, Teushen or Tehuesh†	Argentina

Table 63: South American Indian Language Groups* (continued)

language	location
B. Mosetene: Chimane, Mosetene† [Moseten]	Bolivia
C. Pano-Tacanan	
1. Panoan: Amahuaca [Amawaca], Cashinahua [Cashinawa]	Brazil, Peru
Capanahua [Capanawa], Cashibo, Conibo-Shipibo (Chama, Setebo, Sensi), Marinahua [Marinawa], Marobo, Nocamán, Pano or Pánobo	Peru
Culino (Curina), Jaminahua, Mayoruna or Maruba, Nastanahua, Nixinahua, Parannahua, Poyanahua, Remo, Shaminahua, Tushinahua [Tushinawa], Waninahua or Catoquino, Yahuanahua (Yawanawa), Yumanahua	Brazil
Arazaire [Arasaire], Atsahuaca [Atsawaca] or Chaspa, Yamiaca or Haauñeiri	Peru
Caripuna	Brazil
Chácobo, Pacahuara [Pacawara]	Bolivia
2. Tacanan: Arasa, Cavineña, Chama or Esseejja, Guarizo, Huarayo (Tianinagua), Mabenaro, Maropa or Reyesano, Sapiboca [Sapiboka], Tacana (Araona, Toromona)	Bolivia
D. Yuracare	Bolivia
45. Makú	Venezuela, Brazil
46. Mascoy [Mascoi] or Lengua: Angaité (Sanapá), Kashiká, or Guaná, Lengua or Enslet or Cocoloth (Mascoy)	Paraguay
47. Mataco-Maccá [Macá]	
1. Ashluslay or Chulupí	Paraguay
Chorotí or Solote or Yofuaha	Paraguay, Argentina
Choropí (Suhin, Sotirai), Mataco or Mataguayo (Guisnay, Nocten, Vejoz)	Argentina
2. Enimagá or Cochaboth†	Paraguay
3. Maccá [Macá] or Towothli	Paraguay
48. Movima (27)	Bolivia
49. Munichi [Muniche]	Peru
50. Mura-Matanawí	
A. Bohurá, Mura, Pirahá	Brazil
B. Matanawí†	Brazil
51. Murato or Candoshi or Shapra	Peru
52. Nambikwara [Nambicuara]: Central, Eastern Nambikwara	Brazil
53. Omurano or Mayna†	Peru
54. Otomaco-Taparita†: Otomaco†, Taparita†	Venezuela
55. Pankarurú [Pancararú]	Brazil
56. Puinave-Maku: Makú, Marahan, Querari	Brazil
Puinave	Colombia
57. Puquina†: Pohena or Callahuaya†, Puquina	Bolivia
58. Quechumaran	
A. Aymaran: Aymara, Cauqui or Jaqaru	Bolivia, Peru
B. Quechuan: Almaguero, Inga	Colombia
Ancash, Ayacucho, Cajamarca, Chasutino, Huánuco, Junín, Lamano, Lima, Mayna, Pasco, Ucayali	Peru
Catamarca-La Rioja, Santiago del Estero	Argentina
Cuzco-Bolivian	Peru, Bolivia
Ecuadorian, Quijos, Tena	Ecuador
Tuichi	Bolivia
59. Sabelan: Sabela or Auca or Huarani, Tiwituey	Peru
60. Sáliva-Piaroan: Maco (Macu), Piaroa	Venezuela
Sáliva	Colombia
61. Sec or Sechura or Tallán or Atalán†	Peru
62. Simacu or Itucale or Arucuaya or Urariña	Peru
63. Tarariu (Tarairiu) or Ochukuyana† [Ochukayana]	Brazil
64. Taruma†	Brazil
65. Tikuna [Ticuna] or Tukuna [Tucuna]	Brazil, Colombia
66. Timote†	Venezuela
67. Tiniguan: Pamigua, Tinigua†	Colombia
68. Trumai	Brazil
69. Tucanoan	
A. Western: Amaguaje†, Coto, Piojé	Peru
Coreguaje [Correguaje], Dätuana, Icaguaje†, Macaguaje, Macuna, Sära, (Ömöa, Buhagana), Siona [Sioni], Tama, Tanimuca, Uantia, Yahuna, Yupua	Colombia
Coretu	Brazil
Secoya	Ecuador
B. Eastern: Bara, Erulia (Paneroa, Tsölöa), Karapana [Carapaná], (Möchda), Pamöä, Siana or Chiranga, Tatapuyo, Waiana, Yarutí or Patsoca	Colombia
Desana, Tuyuka (Tuyuca), Wanana (Waikina)	Colombia, Brazil
Kubeo (Cubeo), Tucano	Brazil
70. Tupian	
A. 1. (Tupí-Guaraní): Apiaká† [Apiacá], Awetí, Canoeiro, Kamayurá [Camayura], Kawaíb [Cawahíb] (Pawate, Parintintin, Wirafed) Kayabí (Cayabí), Shetá†, Takuñapé† [Tacunyapé], Tapirapé, Tenetéhara (Anambé, Guayayara [Guajajára], Manajé, Tembé, Turiwara†, Urubú), Tupí-Guaraní (Tupí-nambá), Neengatú	Brazil
Oyampí-Emerillon†	Brazil, Fr. Guiana
Pauserna	Bolivia, Brazil
Guaraní	Argentina, Brazil, Paraguay
Kaiwá	Brazil, Paraguay
Chiriguano, Guarayú	Bolivia
Tapieté, Chané	Paraguay
2. Cocama	Peru
Omagua	Brazil, Peru
3. Guayakí	Paraguay
4. Mawé	Brazil
5. Sirionó	Bolivia
B. Arara, Ramarama (Itogapid), Urukú, Urumí	Brazil
C. Arikem†, Kabishiana†, Karitiana†	Brazil
D. Arué, Digüt, Mondé†	Brazil
E. Guarategaya† (Amniapé†, Kanoe, Mekens), Kepkiriwat†, Makurap†, Tuparí, Wayoró (Apichum)	Brazil
F. Kuruaya† (Curuaya), Mundurukú	Brazil
G. Manitsawá†, Shipaya†, Yuruna	Brazil
H. Puruborá	Brazil
71. Tushá	Brazil
72. Tuyoneiri or Arasairi or Huachipairi	Peru
73. Uman or Huamoi	Brazil
74. Xukurú or Ichikile	Brazil
75. Yabutí†: Aricapú†, Mashubi†, Yabutí or Quipiu†	Brazil
76. Yagua: Masamae, Peba or Nijamo, Yagua or Mishara, Yameo or Camuchivo	Peru
77. Yámana† or Yaghan (41)	Chile
78. Yunca or Chimú or Muchic or Chincha†: Puruhá-Cañari†	Ecuador
Yunca†	Peru
79. Yurí†	Brazil, Colombia
80. Yurimanguí†	Colombia
81. Zamuco: Chamacoco (Ebidoso, Tumrahá), Zamuco (Ayoré, Moro)	Paraguay
82. Záparo: Arabela, Iquito (Cahuarano), Shimigae† (Semigae)†, Andoa, Záparo†	Peru

*Only languages attested linguistically are included. Extinct languages are marked with †. A number in parentheses after the name of a group indicates a possible relationship with the group identified by that number. Languages are separated by commas, names in parentheses are of dialects, and names in brackets are alternative spellings. Except for Arawakan, Macro-Ge, and Tupian, most groupings are geographical, but those identified by capital letters represent in general markedly differentiated groups. Spelling follows the most common usage for each language or group, thus it is not consistent. Equivalent spellings: *b* = *v*; *g* = *j* = *y*; *gu* = *hu* = *u* = *w*; *i* = *y*; *h* = (nothing); *h* = *j*; *k* = *c* (before *a*, *o*, *u*); *k* = *qu* (before *i*, *e*); *sh* = *x* = *ch*; *s* = *z*; *ñ* = *nh* = *ny*; *x* = *j*. Names are arranged alphabetically within each subdivision.

most important group today—Chocó in western Colombia—is distantly related to the rest of the stock. Other languages are Carib in Suriname, Trio in Suriname and Brazil, and Waiwai, Taulipang, and Makushí (Macusí) in Brazil. A relationship with Tupian seems certain.

Tupian. With the exception of Emerillon and Oyampi of French Guiana and northeastern Brazil, Tupian languages were spoken south of the Amazon, from the Andes to the Atlantic Ocean and down to the Río de la Plata. There are approximately 50 attested languages related on the stock level and subdivided into eight families. Tupinambá, the language spoken along the Atlantic coast at the time of discovery, became important in a modified form as a lingua franca, and the closely related Guaraní became the national language in Paraguay, being one of the few Indian languages that does not seem to yield under the influence of Spanish or Portuguese. At the time of discovery, Tupí-Guaraní tribes were moving everywhere south of the Amazon, subjugating other tribes; some of these tribes adopted Tupí-Guaraní. Both Tupí and Guaraní are among the languages that have exerted a great influence on Portuguese and Spanish language. Tupí groups have declined markedly, 26 groups becoming extinct in Brazil between 1900 and 1957, and at least 14 languages disappearing during the same period. The westernmost language, Cocama in Peru, is still spoken by about 19,000 speakers, and Chiriguano in Bolivia has about 20,000 speakers. Other languages have a much smaller number of speakers; there are 19,000 speakers for the 26 surviving groups in Brazil. The total number of Indian speakers of Tupian languages is approximately 60,000, but there are also about 3,000,000 culturally non-Indian speakers of Guaraní in Paraguay. Besides the connection with Cariban, further relationships possibly exist with Macro-Ge, various small families like Zamuco and Mataco-Maccá and isolated languages like Cayuvava.

The Tupí and Guaraní languages

Macro-Ge. Macro-Ge is geographically the most compactly distributed of the big South American language families. Ge proper extends uninterruptedly through inland eastern Brazil almost as far as the Uruguayan border. There are about 10 Ge languages with a total of 2,000 speakers. Most of the other families, now extinct, were located closer to the Atlantic coast, from where they probably were displaced by Tupian expansion. The Bororan family is represented by Bororo in Brazil and by the Otuké language in Bolivia. It seems likely that Macro-Ge has its closest relationship with Tupian.

Quechumaran. Quechumaran, which is composed of the Quechuan and Aymaran families, is the stock with the largest number of speakers—7,000,000 for Quechuan and 1,000,000 for Aymaran—and is found mainly in the Andean highlands extending from southern Colombia to northern Argentina. The languages of this group have also resisted displacement by Spanish, in addition to having gained in numbers of speakers from the time of the Incas to the present as several other groups adopted Quechuan languages. Cuzco-Bolivian Quechua is spoken by well over 1,000,000 speakers, and there are around seven Quechuan languages in Peru with almost 100,000 speakers each. Although most Quechuan languages have been influenced by Spanish, Quechuan in turn is the group that has exerted the most pervasive influence on Spanish. No convincing further genetic relationship has been yet proposed.

Tucanoan. Tucanoan, which is spoken in two compact areas in the western Amazon region (Brazil, Colombia, and Peru), includes about 30 languages with a total of

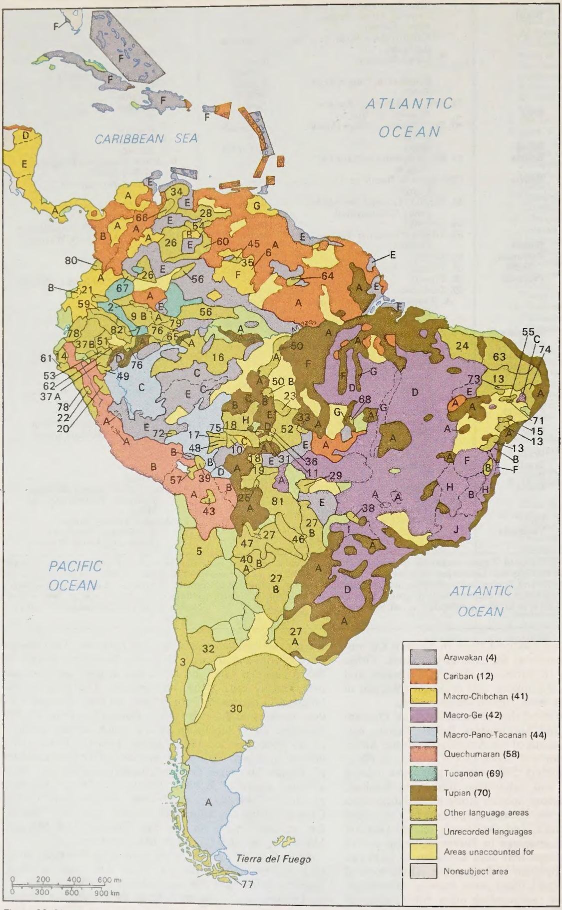


Figure 36: Distribution of the South American Indian languages. The numbers and letters refer to the numbers and letters in Table 62.

over 30,000 speakers. One of the languages is a lingua franca in the region.

Macro-Pano-Tacanan. Macro- Pano-Tacanan, a group more distantly related than a stock, includes about 30 languages, many of them still spoken. The languages are located in two widely separated regions: lowland eastern Peru and adjoining parts of Brazil and lowland western Bolivia on the one hand, and southern Patagonia and Tierra del Fuego on the other. In the latter region the languages are practically extinct.

Other language groups

By number of component languages, or by number of speakers, or by territorial extension, the other language groups are not as significant as those just listed. Most of these small families and isolated languages are located in the lowlands, which form an arch centred on the Amazon from Venezuela to Bolivia and include the bordering parts of Brazil.

Lingua francas and cultural tongues. Lingua francas as well as situations of bilingualism arose mainly under conditions furthered or created by Europeans, although a case like that of the Tucano language, which is used as a lingua franca in the Río Vaupés area among an Indian population belonging to some 20 different linguistic groups, may be independent of those conditions. Quechua, originally spoken in small areas around Cuzco and in central Peru, expanded much under Inca rule, coexisting with local languages or displacing them. It was the official language of the Inca Empire, and groups of Quechua speakers were settled among other language groups, although the language does not seem to have been systematically imposed. The Spaniards, in turn, used Quechua in a great area as a language of evangelization—at one period missionaries were required to know the language—and continued to spread it by means of Quechua speakers who travelled with them in further conquests. During the 17th and 18th centuries it became a literary language in which religious, historical, and dramatic works were written. Today its written literary manifestations are not spontaneous, but there is abundant oral poetry, and in Bolivia radio programs are broadcast entirely in this language.

Dispersion of Tupí-Guaraní dialects, taking place shortly before the arrival of Europeans and even after it, resulted not from imperial expansion—as for Quechua—but from extreme tribal mobility and the cultural and linguistic absorption of other groups. Under Portuguese influence the modified form of Tupinamba known as *língua-geral* ("general language") was the medium of communication between Europeans and Indians and among Indians of different languages in Brazil. It was still in common use along the coast in the 18th century, and it is still spoken in the Amazon. Tupí, now extinct, was an important language of Portuguese evangelization and had a considerable literature in the 17th and 18th centuries. Another dialect, Guaraní, was the language of the Jesuit missions and also had abundant literature until the middle of the 17th century when the Jesuits were expelled and the missions dispersed. Nevertheless, Guaraní survived in Paraguay as the language of a culturally non-Indian population and is today the only Indian language with national, although not official, status—persons not speaking Guaraní being a minority. Paraguayan Guaraní is also a literary language, not so much for learned works—for which Spanish is used—but for those of popular character, especially songs. There is a more or less standardized orthography, and persons literate in Spanish are also literate in Guaraní. A great mutual influence exists between Guaraní and Spanish.

Grammatical diversity

Grammatical characteristics. Diversity rather than common traits characterizes the grammar of South American Indian languages. Features commonly encountered seem to reflect facts of frequency in general typology rather than traits specific to this area. The greatest number of languages are probably suffixing languages like Quechumaran and Huitotoan, or use many suffixes and some prefixes like Arawakan and Panoan. Also very numerous are those languages having few prefixes and suffixes, such as Ge, Carib, or Tupian. Languages employing only prefixes to show grammatical distinctions have not been reported. There are a few with many prefixes but still more suffixes (Jebero, or Chébero); others, like Ona and Tehuelche, with almost no affixing, are also rare.

Similarly, the complexity of words varies a great deal. In Guaraní words with three components and in Piro (Arawakan) words with six elements are of average complexity for the respective languages. In languages like the Cariban or Tupian ones, word roots are nominal (nouns) or verbal (verbs) and may be converted into the other class by derivational affixes; in languages like Quechua or Araucanian, many word roots are both nominal and verbal. Languages like Yuracare form many words by reduplication (the repetition of a word or a part of a word), a process that does not occur systematically in the Tupian languages. Compounding, the joining of two or more words to form new words, is a very widespread type of word formation, but it can be nearly absent, as in the Chon languages. Verb stems in which the nominal (noun) object is incorporated are also rather frequent. Many languages are of the agglutinative type (Quechuan, Panoan, Araucanian); *i.e.*, they combine several elements of distinctive meaning into a single word without changing the element. Others (Cariban, Tupian) show a moderate amount of change and fusion of the elements when combined in words.

Grammatically marked gender in nouns occurs in Guaycuruan (Guaicuruan), and a difference in masculine and feminine gender in the verb occurs in Arawakan, Huitotoan (Witotoan), and Tucanoan, but genderless languages are more common. Singular and plural in the 3rd person ("he, she, it") is not obligatorily distinguished in Tupian and Cariban, but languages like Yámana and Araucanian have singular, dual, and plural. A very common distinction is that between inclusive 1st person ("you and I," hearer included) and exclusive 1st person ("he and I," hearer exluded). Pronominal forms differentiated according to categories that indicate whether the person is present or absent, sitting or standing, and so forth occur in Guaycuruan languages and Movima. Case relations in nouns are generally expressed by suffixes or postpositions; the use of prepositions is rare. Possession is indicated predominantly by prefixes or suffixes, and systems in which possessive forms are the same as those used as the subject of intransitive verbs and as the object of transitive ones are rather common. Classificatory affixes that subclassify nouns according to the shape of the object occur in the Chibchan, Tucanoan, and Waican groups.

Indication of case relationships

Very frequently the verbal forms express the subject, object, and negation in the same word. The categories of tense and aspect seem to be about evenly represented in South American languages, but the specific categories expressed vary a great deal from language to language: Aguaruna (Jívaroan) has a future form and three past forms differentiated as to relative remoteness, while in Guaraní the difference is basically between future and nonfuture. Other languages like Jebero express fundamentally modal categories. Very common are affixes indicating movement, chiefly toward and away from the speaker, and location (*e.g.*, in Quechumaran, Záparo, Itonama), and in some stocks like Arawakan and Panoan there are many suffixes in the verb with very concrete adverbial meaning, such as "by night," "during the day." Classificatory affixes indicating the way the action is performed—by biting, striking, walking—occur in Jebero and Tikuna (Ticuna). Actions done individually or collectively are differentiated paradigmatically in Carib, while in Yámana and Jívaro different verbal stems are used according to whether the subject or the object is singular or plural. There are also various languages (Guaycuruan, Mataco, Cocama) in which some words have different forms according to the sex of the speaker.

Equational sentences are very common. These are formed by juxtaposing two nominal expressions (nouns) without a linking verb, a fact that usually correlates with the absence of a verb "be" for expressing identification or location (*e.g.*, "John good man," "my house there"). Sentences in which the predicate is a noun inflected like a verb with the meaning "being" or "having" that thing designated by the noun also occur in Bororo and Huitoto (Witoto); *e.g.*, "I-knife" = "I have a knife." Sentences in which the subject is the undergoer of the action are frequent, but true passive sentences in which the undergoer and the agent are expressed are rare, though they do occur in Huitoto. Subordinate sentences are rarely introduced by conjunctions; subordination is usually expressed by postposed elements or special forms of the verbs such as gerunds, participles, or subordinate conjugations.

Great variety in distinctive sounds

Phonological characteristics. As in grammar, there are no phonological features common to all South American languages that would be specific to them alone. The number of distinctive sounds (phonemes) may vary from 42 in Jaqaru (Quechumaran) to 17 in Campa (Arawakan). Jaqaru has 36 consonants, while Makushí (Cariban) has 11; some Quechuan languages have only three vowels, whereas Apinayé (Macro-Ge) has ten oral vowels and seven nasal ones. A dialect of Tucano (Tucanoan) exhibits three contrasting points of articulation, while Chipaya (Macro-Mayan) has nine. Many types of contrasting sounds occur although not with equal frequency. Voiceless stops (*e.g.*, *p*, *t*, *k*) occur everywhere, but voiced stops (*e.g.*, *b*, *d*, *g*) may be absent, and fricatives (*e.g.*, *f*, *v*, *s*, *z*) may be few in number. Glottalized voiceless stops—consonants made with simultaneous closure of the glottis and without vibration of the vocal cords—are rather common (Quechumaran, Chibchan), but not glottalized voiced stops (in which the vocal cords vibrate). Also less frequent are aspirated (Quechumaran) and palatalized sounds (Puinave); glottalized nasal sounds (Movima) and voiceless laterals (*l*-like sounds, as in Vilela) are rare. A distinction between velar and postvelar sounds occurs in Quechumaran and Chon, between velar and labiovelar in Tacana and Siona (Sioni); palatal retroflex consonants, made with the tip of the tongue turned up touching the palate, occur in Pano-Tacanan and Chipaya.

Systems with nasal vowels are common (Macro-Ge, Sabelan), but in several languages (Tupian, Waican) nasalization is a feature not of vowels and consonants but of whole words. There is an apparent absence of front rounded vowels (*ü*, *ö*), but central or back unrounded vowels (ɨ, *i*) are common. Systems with long vowels occur in Chipaya and some Cariban languages, and glottalized vowels occur in Tikuna and Chon languages. Very common are pitch-stress systems with high and low tones on stressed syllables; *e.g.*, in Panoan, Huitotoan, and Chibchan. More complex systems with three tones as in Acaricuara, four as in Mundurukú (Mundurucú), and five as in Tikuna are rare. Syllables are generally without complex consonant clusters.

The typology proposed by Tadeusz Milewski, a Polish linguist, classifies American Indian languages into three types: (1) Atlantic, with few oral consonants but complex systems of nasal consonants, and oral and nasal vowels, of which the Ge languages would be typical; (2) Pacific, with complex systems of oral consonants (many contrasting points and modes of articulation) but with few nasal consonants and few vowels, as exemplified by Quechumaran; and (3) Central, with consonant systems more like the Pacific type and vowel systems like the Atlantic, of which Chibcha would be typical. The typology is probably too gross to accommodate meaningfully every language type found in South America, but it holds to a certain extent, especially for the Atlantic type (Macro-Ge, Tupian, and Cariban).

Loan-words from Spanish and Portuguese

Vocabulary. Indian languages vary significantly in the number of loanwords from Spanish and Portuguese. Massive borrowing has taken place in areas where languages have been in intense and continued contact with Spanish or Portuguese, especially where groups are economically dependent on the national life of the country and there is a considerable number of bilingual persons, as in Quechuan, or where no cultural differences correlate with language differences, as in Paraguayan Guaraní. Borrowings have not been limited to designations of artifacts of European origin but affect all spheres of vocabulary, having displaced native terms in many cases. Neither are they limited to lexical items; they include function elements such as prepositions, conjunctions, and derivative suffixes. Sound systems have also been modified. In some contact situations in which the Indian group displayed an antagonistic attitude toward the European conquest, purism developed and loans are comparatively few; *e.g.*, Araucanian. When contact has been frequent but superficial, loanwords are usually scant, but the meaning of native terms has shifted or new descriptive terms have been coined to designate new cultural traits, as in Tehuelche.

Borrowings among Indian languages may have been more numerous than yet reported, judging from the wide and rapid diffusion that loans from Spanish and Portuguese had through the central part of South America. Borrowings between Quechua and Aymara have occurred in great number, but the direction of borrowing is difficult to determine. Many Indian languages in the Andes and the eastern foothills have borrowed from Quechua either directly or through Spanish. In Island Carib (an Arawakan language), borrowings from Carib (a Cariban language) have formed a special part of the vocabulary, properly used only by men; these words were adopted after the Island Carib speakers were subjugated by Caribs.

In turn, some Indian languages have been a source of borrowing into European languages. Taino (Arawakan), the first language with which Spaniards had contact, furnished the most widespread borrowings, including "canoe," "cacique," "maize," and "tobacco," among many others. No other South American Indian language has furnished such widespread and common words, although Quechua has contributed some specialized items such as "condor," "pampa," "vicuña." The larger number of Arawakan borrowings results from these languages having been predominant in the Antilles, a region where Dutch, French, English, Portuguese, and Spanish were present for a long time. Cariban languages, the other important group in that region, do not seem to have furnished many words, but "cannibal" is a semantically and phonetically modified form of the self-designation of the Caribs. The influence of some Indian languages on regional varieties of Spanish and Portuguese has been paramount. Thus Tupí accounts for most Indian words in Brazilian Portuguese, Guaraní in the Spanish of Paraguay and northeast Argentina; and Quechua words are abundant in Spanish from Colombia to Chile and Argentina. In addition, Quechuan and Tupí-Guaraní languages account for most place-names in South America.

Language and culture

No detailed studies are available concerning the relationship of the vocabularies of Indian languages to the culture. Certain areas of vocabulary that are particularly elaborated in a given language may reflect a special focus in the culture, as for example the detailed botanical vocabularies for plants of medical or dietary importance in Quechua, Aymara, and Araucanian. Shifts in cultural habits may also be reflected in the vocabulary, as in Tehuelche, which formerly had a vocabulary designating different kinds of guanaco meat that is now very much reduced, because the group no longer depends on that animal for subsistence. Kinship terminology is usually closely correlated with social organization so that changes in the latter are also reflected in the former: in Tehuelche, former terms referring to paternal and maternal uncles tend to be used indiscriminately, even replaced by Spanish loans, because the difference is not functional in the culture any more.

Proper names, to which different beliefs are attached, offer a variety of phenomena, among them the practice of naming a parent after a child (called teknonymy) in some Arawakan groups; the repeated change of name according to various fixed stages of development, as in Guayaki; word taboo, forbidding either the pronunciation of one's own name or the name of a deceased person, or both, as in the southernmost groups (Alacaluf, Yámana, Chon) and in the Chaco area (Toba, Terena); and the use of totemic names for groups, as in Panoan tribes.

Writing and texts. The existence of pre-Columbian native writing systems in South America is not certain. There are two examples, that of the Cuna in Colombia and an Andean system in Bolivia and Peru, but in both cases European influence may be suspected. They are mnemonic aids—a mixture of ideograms and pictographs—for reciting religious texts in Quechua and ritual medical texts in Cuna. The Cuna system is still in use.

Although the linguistic activity of missionaries was enormous and their work, from a lexicographic and grammatical viewpoint, very important, they failed to record texts reflecting the native culture. The texts they left for most languages are, with a few exceptions, of a religious nature. Most of the folklore has been collected in the 20th century, but many important collections (*e.g.*, for the Fuegian and Tacanan tribes) are not published in the native language but rather in translation. There are good texts recorded in the native language for Araucanian, Panoan, and Cuna, for instance, and more are being recorded by linguists now, though not necessarily analyzed from a linguistic point of view.

Efforts are being made in several areas to introduce literacy in the native Indian languages. For some, practical orthographies have existed since the 17th century (Guaraní, Quechua); for several others, linguists have devised practical writing systems and prepared primers in recent years. The success of these efforts cannot yet be evaluated. (J.A.S.)

LANGUAGE ISOLATES

Sumerian language

Sumerian is the oldest written language in existence. First attested about 3100 BC in southern Mesopotamia, it flourished during the 3rd millennium BC. About 2000 BC, Sumerian was replaced as a spoken language by Semitic Akkadian (Assyro-Babylonian) but continued in written usage almost to the end of the life of the Akkadian language, around the beginning of the Christian era. Sumerian never extended much beyond its original boundaries in southern Mesopotamia; the small number of its native speakers was entirely out of proportion to the tremendous importance and influence Sumerian exercised on the development of the Mesopotamian and other ancient civilizations in all their stages.

History. Four periods of Sumerian can be distinguished: Archaic Sumerian, Old or Classical Sumerian, New Sumerian, and Post-Sumerian.

Archaic Sumerian covered a period from about 3100 BC, when the first Sumerian records make their appearance, down to about 2500 BC. The earliest Sumerian writing is almost exclusively represented by texts of business and administrative character. There are also school texts in the form of simple exercises in writing signs and words. The Archaic Sumerian language is still very poorly understood, partly because of the difficulties surrounding the reading and interpretation of early Sumerian writing and partly because of the meagreness of sources.

Old, or Classical, Sumerian

The Old, or Classical, period of Sumerian lasted from about 2500 to 2300 BC and is represented mainly by records of the early rulers of Lagash. The records are business, legal, and administrative texts, as well as royal and private inscriptions, mostly of votive character; letters, both private and official; and incantations. These sources are much more numerous than those of the preceding period, and the writing is explicit enough to make possible an adequate reconstruction of Sumerian grammar and vocabulary.

During the period of the Sargonic dynasty, the Semitic Akkadians took over the political hegemony of Babylonia, marking a definite setback in the progress of the Sumerian language. At this time the Akkadian language was used extensively throughout the entire area of the Akkadian empire, while the use of Sumerian gradually was limited to a small area in Sumer proper. After a brief revival during the 3rd dynasty of Ur, the New Sumerian period came to an end about 2000 BC, when new inroads of the Semitic peoples from the desert succeeded in destroying the 3rd dynasty of Ur and in establishing the Semitic dynasties of Isin, Larsa, and Babylon.

The period of the dynasties of Isin, Larsa, and Babylon is called the Old Babylonian period, after Babylon, which became the capital and the most important city in the country. During this time the Sumerians lost their political identity, and Sumerian gradually disappeared as a spoken language. It did, however, continue to be written to the very end of the use of cuneiform writing. This is the last stage of the Sumerian language, called Post-Sumerian.

In the early stages of the Post-Sumerian period the use of written Sumerian is extensively attested in legal and administrative texts, as well as in royal inscriptions, which are often bilingual, in Sumerian and Babylonian. Many Sumerian literary compositions, which came down from the older Sumerian periods by way of oral tradition, were recorded in writing for the first time in the Old Babylonian period. Many more were copied by industrious scribes from originals now lost. The rich Sumerian literature is represented by texts of varied nature, such as myths and epics, hymns and lamentations, rituals and incantations, and proverbs and the so-called wisdom compositions. For many centuries after the Old Babylonian period, the study of Sumerian continued in the Babylonian schools. As late as the 7th century BC, Ashurbanipal, one of the last rulers of Assyria, boasted of being able to read the difficult Sumerian language, and from an even later period, in Hellenistic times, there are some cuneiform tablets that show Sumerian words transcribed in Greek letters.

Figure 37: Sumerian inscription, detail of a diorite statue of Gudea of Lagash, 22nd century BC. In the Louvre, Paris.
Archives Photographiques

Rediscovery. Around the time of Christ, all knowledge of the Sumerian language disappeared along with that of cuneiform writing, and in the succeeding centuries even the name Sumer vanished from memory.

Unlike Assyria, Babylonia, and Egypt, whose histories and traditions are amply documented in biblical and classical sources, there was nothing to be found in non-Mesopotamian sources to make one even suspect the existence of the Sumerians in antiquity, let alone fully appreciate their important role in the history of early civilizations.

Decipherment of Sumerian cuneiform

When the decipherment of cuneiform writing was achieved in the early decades of the 19th century, three languages written in cuneiform were discovered: Semitic Babylonian, Indo-European Persian, and Elamite, of unknown linguistic affiliation. Only after the texts written in Babylonian had become better understood did scholars become aware of the existence of texts written in a language different from Babylonian. When the new language was discovered it was variously designated as Scythian, or even Akkadian (that is, by the very name now given to the Semitic language spoken in Babylonia and Assyria). It was only after knowledge of the new language had grown that it was given the correct name of Sumerian.

Characteristics. The linguistic affinity of Sumerian has not yet been successfully established. Ural-Altaic (which includes Turkish), Dravidian, Brahui, Bantu, and many other groups of languages have been compared with Sumerian, but no theory has gained common acceptance. Sumerian is clearly an agglutinative language in that it preserves the word root intact while expressing various grammatical changes by adding on prefixes, infixes, and suffixes. The difference between nouns and verbs, as it exists in the Indo-European or Semitic languages, is unknown to Sumerian. The word *dug* alone means both "speech" and "to speak" in Sumerian, the difference between the noun and the verb being indicated by the syntax and by different affixes.

The distinctive sounds (phonemes) of Sumerian consisted of four vowels, *a, i, e, u,* and 16 consonants, *b, d, g,* ŋ, *h, k, l, m, n, p, r, s, ś, š, t, z.* In Classical Sumerian, the contrast between the consonants *b, d, g, z* and *p, t, k, s* was not between voiced (with vibrating vocal cords) and voiceless consonants (without vibrating vocal cords) but between consonants that were indifferent as to voice and those that were aspirated (pronounced with an accompanying audible puff of breath). The semivowels *y* and *w* functioned as vocalic glides.

In the noun, gender was not expressed. Plural number was indicated either by the suffixes *-me* (or *-me* + *esh*), *-hia*, and *-ene*, or by reduplication, as in *kur* + *kur* "mountains." The relational forms of the noun, corresponding approximately to the cases of the Latin declension, include: *-e* for the subject (nominative), *-a(k)* "of" (genitive), *-ra* and *-sh(e)* "to," "for" (dative), *-a* "in" (locative), *-ta* "from" (ablative), *-da* "with" (commitative).

The Sumerian verb

The Sumerian verb, with its concatenation of various prefixes, infixes, and suffixes, presents a very complicated picture. The elements connected with the verb follow a rigid order: modal elements, tempo elements, relational elements, causative elements, object elements, verbal root, subject elements, and intransitive present–future elements. In the preterite transitive active form, the order of object and subject elements is reversed. The verb can distinguish, in addition to person and number, transitivity and intransitivity, active and passive voice, and two tenses, present-future and preterite.

Several Sumerian dialects are known. Of these the most important are *eme-gir*, the official dialect of Sumerian, and *eme-SAL*, the dialect used often in the composition of hymns and incantations (see also WRITING).

(I.J.G.)

Etruscan language

The Etruscan language was spoken by close neighbours of the ancient Romans. The Romans called them Etrusci or Tusci; in Greek they were called Tyrsenoi or Tyrrhenoi; in Umbrian, and Italic language, their name can be found in the adjective *turskum*. The Etruscans' name for themselves was *rasna* or *raśna*.

The Etruscans lived in Italy in the region of modern Tuscany, in an area bounded by the Arno River on the north, the Tiber River on the southeast, and the Tyrrhenian Sea on the west. At one time they controlled most of an area extending south from Milan through Marzabotto and Sarsina to the Adriatic Sea north of Ancona, and to the southwest their rule extended as far as Capua, Naples, and Pompeii. For the history of the Etruscans and Etruria, see GRECO-ROMAN CIVILIZATION: *Ancient Italic peoples.*

Records and scholarship. The Etruscan language is known mainly from epigraphic records originating in the Tuscan area and dating from the 7th century BC to the first years of the Christian Era.

Inscriptions

There are some 10,000 of these inscriptions, mainly brief and repetitious epitaphs or dedicatory formulas, as well as votive or owner's inscriptions on paintings in tombs and accompanying engraved figures on small artifacts such as metal mirrors. There are, however, some remarkable exceptions to the general brevity of the inscriptions, and there are important differences in their origins. The longest single text, of 281 lines (about 1,300 words), now in the National Museum at Zagreb, is written on a roll of linen that had been cut into strips and used in Egypt as a wrapping for a mummy; a clay tablet found at Capua contains some 250 words; a stone slab from Perugia has two adjacent sides elegantly engraved with an inscription of 46 lines (some 125 words); a bronze model of a liver found at Piacenza, which probably represents the Etruscan microcosm in a form used for instruction in divination, has some 45 words; and a heavy rectangular block found on the island of Lemnos in the northern Aegean has an engraving of what is probably a warrior with one inscription of perhaps 18 words surrounding the head and another of 16 words in three lines on an adjacent side. In 1964 two inscriptions on gold tablets, one in Phoenician and the other in Etruscan, were unearthed at Pyrgi.

In the Museo Archeologico Nazionale dell'Umbria, Perugia, Italy

Figure 38: Etruscan inscription from a section of the "Cippus of Perugia."

Despite many attempts at decipherment and some claims of success, the Etruscan records still defy translation. While the possibility always remains that an imaginative conjecture or a brilliant inference will suddenly provide the key to the mystery, this now seems remote. The etymological method of investigation, which ultimately depends upon the recognition of presumed cognates from related languages, seems to have failed because no clear and certain relationship between Etruscan and any other language has ever been established. The procedure sometimes called the combinatory method now appears to be the most efficacious if not indeed the only useful one. It requires, first, that note be made of anything unusual in the provenance of the object on which Etruscan writing is found (such as that the mummy wrapping came from Egypt and the Lemnos inscription from the Aegean) and likewise of anything unusual in the object itself (*e.g.,* that it is a bronze replica of a liver or the representation of a god or mythological figure). Finally, each word and phrase and formula is compared with every recurrence of the same element or elements elsewhere, and all variations in the physical and the linguistic contexts are recorded. By this means it has been possible to assign some words to grammatical categories such as noun and verb, to identify some inflectional endings, and to assign meanings to a few words of very frequent occurrence.

The problem of Etruscan origins is insoluble until the language can be translated. While nothing at all is certain other than the existence of Etruscans in Italy, some Etruscan writing in Egypt, and an Etruscan inscription on Lemnos, the weight of all the evidence seems to favour a non-Italic but certainly Mediterranean place of origin of the Etruscan people.

Linguistic relationships

It is unlikely, therefore, that the Etruscan language is genetically related to any language or language family existing in an area remote from the Mediterranean. On the other hand, it does not follow that Etruscan must be related to a language or language fragment that can be found in the Mediterranean area.

The Etruscan alphabet. Etruscan is written in an alphabet probably derived from one of the Greek alphabets.

It is of very great importance that Etruscan is written in a recognizable alphabet related to the Greek and Semitic because sound values can be assigned with some degree of precision to each symbol. Etruscan writing proceeded from right to left and in earliest times had no word division or punctuation. In about the 6th century BC a system of points, or dots, consisting of four, three, or two dots inscribed vertically, was introduced to mark word boundaries and, in some instances, apparently, to indicate syllables and possibly abbreviations.

By courtesy of the Soprintendenza alle Antichita, Florence

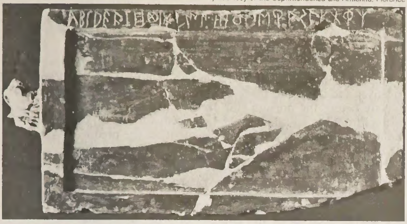


Figure 39: Etruscan alphabet on the edge of a writing tablet from Marsiliana d'Albegna, 7th century. In the Museo Archeologico, Florence.

There were four vowels in Etruscan, *i, e, a,* and *u* or *o,* and symbols in the alphabet for *p, t, c, m, n, l, r, z* and for the equivalents of the Greek *phi, theta,* and *chi,* which in Etruscan as in classical Greek were the aspirated stops *ph, th, ch* (pronounced as *p, t, k* with an added brief puff of air). There were two sibilants, written *s* and *ś,* for which the precise pronunciation is uncertain; two front fricatives, *f* and *v,* articulated either with the two lips (bilabial) or with the lower lip approaching the upper front teeth (labiodental); and an *h,* which nearly always occurs at the beginning of words and is used to represent, inconsistently, the rough breathing of Greek (*e.g.,* Greek *Hēraklēs,* Etruscan *hercle* or *ercle*). There were also a *k* and a *q,* of which the precise pronunciation is unknown. A marked tendency to make all vowels in a word similar or identical (qualitative vowel harmony) is characteristic: Greek *Klutaimēstra,* which if transliterated directly into Etruscan would be *cluthemestha,* actually occurs as *cluthumustha* and *clutmsta.*

Both historical changes and dialectal differences can be observed. Diphthongs became single letters. Thus Greek *Aiwas* became Etruscan *aivas, eivas,* and *evas,* successively; *au* alternated with *a*; *eu* (like *ai*) became *e* (Greek *Kleopatra* is Etruscan *clepatra;* Greek *Poludeukēs* is, with Etruscan vowel harmony, *Pultuce*). Among consonants the most noticeable changes are *c* to *ch* to *h* (*e.g., casri* becomes *chasri, caspr* becomes *haspr*); similarly, *p* changes to *ph* to *f* to *h* and *t* to *th* to *h.* Throughout the history of Etruscan, a first syllable usually remains unchanged, whereas later syllables tend to weaken or lose vowels, at least in writing. Older Etruscan *lavtun* "family" becomes in later Etruscan *lavtn;* other examples are *mutana* changing to *mutna,* Greek *Adonis* written *atunis* and then *atuns,* Greek *Alexandros* appearing as *elchsntre.* The consonant cluster of *elchsntre,* while extreme, is not untypical of Etruscan spelling; words thus written have led some to suggest that a very economical spelling system may have been used that was far removed from the reality of pronunciation, requiring the introduction of lightly stressed vowels in actual utterance. (For a short history and a chart of the Etruscan alphabet, see WRITING: *Alphabetic writing.*)

Grammatical characteristics. The minimal unit of meaning seems to have been a verbal root, such as *zic* or *zich,* meaning "write." The suffixing of any vowel or certain consonants (*c* or its variant *ch, t* or its variant *th, l, r,* or *n*) produced a noun. The vowel *u* was used to form a gerund that, without further change, could be used as an agent noun; thus *zicu* meant "writing," then, further, "scribe," and then the equivalent of the Roman name Scribonius. From these verbal nouns, denominative verbs could be made; thus from *zicu* plus *-ce,* a past tense of perfective affix, was made *zichuche* "he wrote, he has written."

Because of the large number of names occurring in the inscriptions, the noun declension system can be understood reasonably well. Similar to the process of word building is the construction called *genitivus genitivi,* or "genitive of the genitive," in which several possessive suffixes may be added to a word in succession. Thus, the simple genitive of *larth,* a proper name, is *larthal* "Larth's"; a second genitive suffix added gives *larthalisa* "of that which is Larth's."

The *genitivus genitivi*

There is apparently no grammatical gender in Etruscan. In late Etruscan an *-i* suffix marks some women's names; still later, apparently (or possibly in a different dialect area), *-ia* is similarly used. Although these suffixes appear to be the same as the final elements in some words designating exclusively female functions, such as *puia* "wife" (in one occurrence spelled *pui,* if it is the same word) and *ati* "mother" (in one occurrence *atiu,* if it is the same word), there is no evidence for any syntactic use of gender, and there is no formal marker that can be shown to have marked gender consistently.

Case endings do not differ from singular to plural; in the singular they are suffixed directly to the word stem, and in the plural they are added to the stem, along with one of the plural markers *ar, er, ur.* There is no distinctive nominative (subject) case marker, the word stem or, in some cases, the root alone serving as the nominative. A final marker *-s,* however, does appear to have been added in some instances of a probable nominative case.

The repetitive nature of most Etruscan inscriptions is such that very few distinctively different verb forms are available for analysis. Indeed, probably the only really certain verbal suffix is *-ce.* It must not be assumed, however, that the paucity of the verbal data from inscriptions reflects an impoverished verb system in the language; indeed, judging from the variety of verbal stems to which the recurring *-ce* is added, it is more likely that the Etruscan verb had a more complicated structure than the noun.

Vocabulary. Since the language is undeciphered, meaning can be assigned with certainty to only a few Etruscan words that occur very frequently in the texts. Some kinship terms are sure—among these are *ati,* "mother," *clan* "son," *śec* "daughter," *puia* "wife." Less certain but probably correct are words designating members of the larger societal organization: *lavtn* "family," *zilc* "official," *maru* "official," *spur* "city," *rasna* or *raśna* "Etruscan, Etruria." A pair of dice certainly have on them the names of the numbers from one through six. Although the order of these numbers has been and still is disputed, the arrangement most generally accepted is this: *thu* "one," *zal* "two," *ci* "three," *śa* "four," *mach* "five," *huth* "six."

Among the continuing mysteries of Etruscan are the reasons why the Etruscans left no written records of their great civilization other than inscriptions and occasional texts and why the Romans, who knew the Etruscans intimately, transmitted little or nothing to posterity about either Etruscan literature or their language, which must certainly have been spoken, or at least preserved, by some families in Rome long after the period of Etruscan greatness had passed. (M.Fo.)

Basque language

Basque, the only remnant of the languages spoken in southwestern Europe before the region was romanized, is currently used in a narrow area of approximately 10,000 square kilometres (3,900 square miles) in Spain and France. The number of Basque-speaking persons outside that territory, in Europe and in the Americas, however, is far from insignificant. In Spain the Basque-speaking region comprises the province of Guipúzcoa, parts of Vizcaya and Navarra, and a corner of Álava, and in France the western region of the *département* of Pyrénées-Atlantiques. Although few statistics are available, the number of speakers, who are largely bilingual, might be judiciously estimated at 1,000,000. Most of them live in the highly industrialized Spanish part of the Basque country.

The Basques have derived their name, Euskaldunak, from Euskara, the native word for their language. According to the classification of the 19th-century philologist Prince Louis-Lucien Bonaparte, there are eight modern dialects of Basque. Dialectal division is not strong enough to mask the common origin or to preclude mutual understanding. Basque attained official status for a short period (1936–37) during the Spanish Civil War, under Basque autonomous government. In 1978, Basque and Castilian Spanish became the official languages of the autonomous Basque Country, which includes Guipúzcoa, Vizcaya, and Álava provinces of Spain.

Origins and classification. Basque remains an isolated language with no known linguistic relatives. The hypothesis of the German philologist Hugo Schuchardt (1842–1927), which once had wide currency, posited an intimate genetic connection between Basque and Iberian (see below) and the Hamito-Semitic (Afro-Asiatic) language group. This theory was superseded by attempts to establish a more or less close link between Basque and Caucasian, the language group indigenous to the Caucasus region. A lack of common linguistic characteristics between the Basque and Hamito-Semitic languages makes Schuchardt's hypothesis extremely dubious. There are, however, some common features that favour the relationship between Basque and Caucasian. Still, proof of a genetic relationship beyond reasonable doubt appears remote. Perhaps the most promising theory involves the comparison of Basque with the long-extinct Iberian, the language of the ancient inscriptions of eastern Spain and of the Mediterranean coast of France. But, despite amazing phonological coincidences, Basque has so far contributed next to nothing to the understanding of the now-readable Iberian texts. Therefore, it is possible that the similarity may have resulted from close contact between Basques and Iberians and not from a genetic linguistic relationship.

Link between Basque and Iberian

History of the language. At the beginning of the Christian Era, dialects of Euskarian (Basque) stock were probably spoken north and south of the Pyrenees and as far east as the Valle de Arán in northeastern Spain. It is likely that only the disruption of Roman administration in these regions saved the Basque dialects from being completely overcome by Latin. It is also likely that the Basque tongue, which had a firm foothold in the country that then began to be called Vasconia, experienced a substantial expansion toward the southwest, which carried it to the Rioja Alta (High Rioja) region in Old Castile and near Burgos. The more eastern Basque dialects, separated from the main area by Romance-speaking populations, were doomed. During the Middle Ages, Basque, the language of a population more peasant than urban, could not possibly hold the field as a written language against Latin and its successors, Navarrese Romance and, to a certain extent, Occitan (the *langue d'Oc*, also called Provençal) in the kingdom of Navarre. Since the 10th century, Basque has slowly but steadily lost ground to Castilian Spanish; in the north, however, where French is a more modern rival, the Basque-speaking area is practically the same as it was in the 16th century. In the last two centuries, above all in industrial centres, Basque has had to fight for survival in the heart of the Basque-speaking country, as well as on the frontier of the Basque-speaking area.

Basque records and writing

Latin inscriptions from the Roman period, found mostly in southwestern France, record a handful of proper names of unmistakable Basque etymology. From AD 1000 on, records consisting chiefly of proper names but also of Basque phrases and sentences grew more numerous and reliable. The first printed Basque book, dating from 1545, began an uninterrupted written tradition. Scholarly Basque literature, with its prevailing religious interests, has been neither abundant nor varied until recent times. Intense efforts are now being made to introduce Basque as a vehicle of private primary education. In addition, a model of a unified, standard written language also seems to be gaining increasing acceptance.

Phonology. The sound pattern of Basque is, on the whole, similar to that of Spanish. The number of distinctive sounds is relatively low compared with other languages. Combinations of sound (*e.g.*, consonant clusters) are subject to severe constraints. It can confidently be asserted that certain types of consonant clusters, such as *tr, pl, dr,* and *bl,* were all but unknown about two millennia ago. The common sound system underlying the systems of the present Basque dialects has five (pure) vowels and two series of stopped consonants—one voiced (without complete stoppage in many contexts), represented by *b, d, g,* and the other voiceless, represented by *p, t, k.* Nasal sounds include *m, n,* and palatal *ñ,* similar to the sound indicated by *ny* in "canyon." In this respect, as in others, Basque orthography coincides with the Spanish norm. There are two varieties of *l,* the common lateral *l* and a palatal variety, *ll,* as in Spanish, that sounds similar to the *lli* in "million" (as *l* + *y*). The Basque *r,* made by a single tap of the tongue against the roof of the mouth, contrasts with a rolled or trilled *r,* written *rr.* Two phonological features are worthy of special attention. Sibilants (both fricatives and affricates) made with the area of the tongue directly before the dorsum (the back of the tongue) are distinct from the apical sibilants, produced with the tip of the tongue. The letter *z* in Basque symbolizes the predorsal fricative, and *tz,* the predorsal affricate sound; *s* and *ts* represent the apical fricative (similar to Castillian Spanish *s*) and affricate, respectively. (A fricative is a sound, such as English *f* or *s,* produced with friction and, hence, without complete stoppage in the vocal tract; an affricate is a sound, such as *ch* in "church" or *j* in "jam," that begins as a stop and ends as a fricative, with incomplete stoppage.) In addition to these hissing sibilants, Basque also includes the hushing ones, written as *x* and *tx*; they are like the English *sh* and *ch.* The *x* and *tx* sounds, along with the palatal sounds written as *ll* and *ñ,* often have an expressive value (diminutive, endearing) in comparison with their nonpalatal counterparts; *e.g., hezur* means "bone" and *hexur* "little bone" (fish bone, for example); *sagu* is "mouse" and *xagu* "little mouse."

Varieties of sibilants

The phonology of some Basque dialects may be more complex than that presented in the preceding paragraph. In the easternmost Souletin region, for example, the dialect has acquired, by internal development or by contact with other languages, a sixth oral vowel—rounded *e* or *i*—and nasal vowels, voiced sibilants, and voiceless aspirated stops. The aspiration accompanying stop consonants consists of a small puff of air. There is also, word-initially and between vowels, an aspirated *h,* once common but now peculiar to the northern dialects. It has also been retained in the proposed standard form of Basque.

Grammar. The mention of two features is unavoidable in describing Basque syntax. Basque is, in the first place, a language of the so-called ergative type. That is, it has a case denoting the agent of an action. Hence, what in English would stand for the subject of a transitive verb is expressed in Basque by means of a suffix *-k*; for example, in the sentence "the foot serves the hand, and the hand serves the foot," *oinak zerbitzatzen du eskua, eta eskuak oina,* the first word, meaning "the foot," is composed of three elements, *oin* "foot," *-a,* "the," and *-k,* which marks the Basque equivalent of the subject of the verb. The fourth word, meaning "the hand," does not have the *-k* ending. In the second clause, *eta eskuak oina,* the word for hand, *eskuak,* now has the ergative *-k* ending to indicate that the hand is the agent of the clause "the hand serves the foot." The subject of an intransitive verb, which is not distinguished from the object of a transitive verb, has no overt mark—*e.g.,* in "if the belly does not eat, the belly itself will fail," *sabelak jaten ez ba du, sabela bera ihartuko da,* the first term, *sabelak* "the belly," has the *-k* marker because it is the agent of a transitive verb "eat"; but, in the second clause, *sabela* is the subject of the intransitive verb "fail" and, therefore, has no overt grammatical mark.

The second characteristic feature of Basque concerns the finite verb, which acts as a summary of all the noun phrases in the sentence. It has markers for all three persons—the 1st, 2nd, and 3rd—and may contain as many as three personal references (for subject, direct object, and indirect object). *Da,* for example, means "is," *du* means "he has it," and *dio* means "he has it for him" in the sentence *oinari ez dio eskuak kolperik emaiten* "the hand does not give a blow [*kolpe*] to the foot [*oin-a-ri*]." In certain situations the interlocutor also can be referred to within the verb. Further, most Basque verbs have only a compound conjugation—*e.g.*, *erori da* "he has fallen," literally, "he is fallen," and *jaten du* "he eats [is eating] it."

Although some ancient prefixes are still apparent in modern Basque, they are no longer productive, so that Basque can be characterized as an overall suffixing language; that is, it appends suffixes to words. There is one declension with suffixes or postpositions to indicate number and case—*e.g.*, *etxe-a* "the house," but *etxe berri-a* "the new house," and *etxe berri-a-ri* "to [for] the new house." Suffixes, under certain restrictions, may be heaped upon one another. Theoretically, genitival endings indicating possession may be added to one another without limit. This is similar to the case in English of "the button of the coat of the son of the Major of York"; in Basque, however, the phrase "of the" is indicated by an ending, *-(r)en*, added to the noun. Noun suffixes also can be attached to verb forms in order to express subordination of the clauses in which the verb forms appear—*e.g.*, *da* "is," *den* "which is," *dena* "that (*-a*) which is," *denean* "when there is," literally, "in that which is." Prefixes also are used for that purpose; *e.g.*, *ez du jaten* "he does not eat" with the particle *ba* "if" becomes "if [the belly] does not eat," *jaten ez ba du*.

Vocabulary. Basque has preserved a peculiar and distinctive appearance, despite the overwhelming pressure to which it has been subjected over a period of at least 2,000 years. Nevertheless, its borrowings from the neighbouring languages, especially of words and idioms, can hardly be underrated. Loanwords from the Romance languages are numerous. Some of them bear the unmistakable stamp of their archaic Latin ancestry—*e.g.*, *bake* "peace" from Latin *pax, pacis*, *bike* "pitch" from Latin *pix, picis*, and *errege* "king" from Latin *rex, regis*. Contrary to a widely held opinion, Indo-European loanwords of non-Latin origin are extremely scarce. Derivation, the formation of new words by the use of suffixes, is accomplished partly through the use of borrowed suffixes. This practice, as well as the compounding of nouns to form new words, as in *bizkar-hezur* "backbone," has been very much alive throughout the history of the language. On the other hand, Basque itself has contributed but little vocabulary to the Spanish, Occitan, French, and English languages. But family and place names of Basque coinage are frequently encountered in Spain and in Latin America, where they can be found in such proper names as Aramburu, Bolívar, Echeverría, and Guevara.

Latin loanwords

(L.M.)

PIDGIN

The term pidgin is applied to a number of varieties of speech that have grown out of English or other languages and that have been used in various parts of the world since the 17th century. Often termed "bastard jargons," "mongrel lingos," or the like, these tongues in fact are languages like any others and can be accurately delimited and described.

Definitions of lingua franca, pidgin, and creole. When a language is used as a means of communication between persons having no other language in common (*e.g.*, French in 18th-century diplomacy), it is a lingua franca. A lingua franca native to none of those using it and with a sharply reduced grammar and vocabulary is called a pidgin. (This definition of pidgin excludes both the broken English of a beginning learner and the skillful but nonnative use of English in such countries as India.) When a whole speech community gives up its former language or languages and takes a pidgin as its mother tongue, the pidgin becomes a creole (is creolized).

Situations in which pidgins developed

Origins. A number of pidgins and creoles have arisen on the basis of various European languages. The first known pidgin, Lingua Franca, or "Westerners' language," of the medieval Levant and the Barbary Coast, was based chiefly on Italian. The American Indians first encountered by Englishmen in the 17th century were a tribe known as Pidians near the mouth of the Orinoco; the reduced language that emerged was termed Pidgin (= Pidian) English. Later in the same century, other varieties of Pidgin English grew up in China as a result of English commercial contacts and in Africa in connection with slave-trading activities. (Some authorities derive the word "pidgin" from a variation of English "business.") Establishment of plantation economies in the Caribbean area, with large groups of Negro slaves from different language backgrounds in West Africa, led to a number of pidgins based on English, French, Spanish, and Portuguese. Many have survived as creoles—*e.g.*, Gullah off the Sea Islands of South Carolina, the English of the Antilles, and Sranan (formerly called Taki-Taki) in Suriname (formerly Dutch Guiana), all based on English; the French-based creoles of Louisiana, Haiti, and the Lesser Antilles; and the Papiamento of Curaçao, an outgrowth of Pidgin Spanish and Portuguese. Early contacts between settlers and natives led to the formation of pidgins in Australia and New Zealand, whereas the Pidgin English of the South Seas (called Beach-la-Mar) grew out of whaling, trading, and recruiting native labour. Pidgin English is extinct in New Zealand and the Caroline Islands and moribund in Australia but still flourishes in Melanesia (Vanuatu, Solomon Islands, New Guinea) and has become creolized in Hawaii. One variety, that of the Rabaul area (New Britain), has become the widespread and indispensable lingua franca of Papua New Guinea because of the official sanction that was given to its use under German rule (1884–1914) and later Australian administrations.

Development and persistence of pidgins

Survival. Pidgin languages spring from the initial, nonintimate contacts between speakers of different languages, when quick comprehension is more highly valued than grammatical correctness or fine shades of meaning. As contacts grow closer, normally one group learns the other's language more fully, and pidgins survive the stages of initial contact only in special circumstances. Pidgins persist where a dominant group regards another as childlike or capable of only a simplified version of the "superior" language, as in the relations between Europeans and American Indians, West Africans, or South Sea natives. On plantations and in other situations where European masters were in permanent contact with native servants or labourers, pidgins served as status languages, as in New Guinea. Caste distinction, however, is not a necessary function of a pidgin; Russonorsk, for instance, was a reduced language used by Russians and Norwegians in the Arctic at the beginning of the 20th century. Chinese Pidgin English survived for three centuries and not only in master–servant relations; it also was in use between English merchants and Chinese dignitaries, primarily because each side desired to keep the other at arm's length. Slaves on Caribbean plantations, New Guinea natives in newly founded multilingual villages, and others who have come to live together with no language in common save a pidgin have used the pidgin as the customary language of the group. In such instances, the resultant creole has usually re-expanded its structure and vocabulary by borrowing from the language of a culturally dominant group—*e.g.*, Haitian Creole from French, Sranan and Papiamento from Dutch, and Melanesian Pidgin from English.

Spelling and sounds

Orthographies. In its original function as a lingua franca among unlettered folk, a pidgin language is a medium of purely oral communication, as also are creoles in their initial stages. Only afterward, and usually in connection with missionary or other educational programs, are spelling systems devised for pidgins or creoles. Speakers of European languages have often applied the orthographic conventions of their own languages, as when the Melanesian Pidgin sentence for "Why did you hit this policeman?" is written *Belong what name you fight 'im dis fellow police boy?* Such a spelling embodies all the inconsistencies of English orthography and is therefore difficult to learn; it distorts the structure of pidgin; and it confirms the naïve European or American in his belief that pidgin is only a ridiculous reduction of English. Those who have devised orthographies, for ease of learning, accuracy in representing linguistic structure, and emphasis on the independent status of the language, have used phonemically based systems. The most effective orthographies of this type use the letters available on typewriters or in printshops, but consistently and predictably. Thus, the Melanesian Pidgin sentence just quoted reads, in the officially recognized orthography: *Bilong wonem yu faitim dispela plisboi*? In the following discussion, forms are cited only in this type of transcription.

Phonology. The simplification which characterizes pidgin extends to all aspects of linguistic structure (sounds, forms, constructions) as well as vocabulary. In some varieties stress is automatically on the first syllable; *e.g.*, *bíkos* "because," *míshin* "machine." A minimum of five distinctive vowels is necessary, those represented in Latin or Italian pronunciations by *a* ("ah"), *e* ("eh"), *i* ("ee"), *o* ("oh"), *u* ("oo"), as in *antap* "up," *em* "he," *winim* "defeat," *kot* "coat," *tu* "also." The vowel sound *a* ("ah") combines with *i* ("ee") and *u* ("oo") in the basic diphthongs *ai* (English "i") and *au* ("ow"), as in *dai* "cease" and *nau* "now." Some, but not all, speakers make further distinctions—*e.g.*, between the *e* of *em* "he" and the *ei* ("ay") of *neim* "name." In almost all varieties of pidgin English, the two consonant sounds represented by English *th* have merged with *t* and *d* respectively: *saut* "south," *dispela* "this." Many speakers of Melanesian Pidgin merge *f* and *p* (in current official orthography, both are represented by *p*) and also merge *ch* and *sh* with *s*: *tumas* "very" (from English "too much"), *masin* "machine." Users of pidgin often carry over habits of sound production from their native languages; *e.g.*, many Melanesian languages have *mb*, *nd* as variants of *b*, *d* between vowels, and thus Melanesians often pronounce *tabak* "tobacco" as *tambak*, and *sidaun* "sit" as *sindaun*.

Morphology. Grammatical categories—such as number, gender, case, person, tense, mood, voice—are almost absent from pidgin and creole languages, as from many other languages of the world. Pidgin is not, however, "devoid of grammar," as is often asserted. Melanesian Pidgin, for example, has three inflected parts of speech: pronouns, adjectives, and verbs. The suffix *-pela* added to pronouns makes plurals, in *mi* ("I, me") versus *mipela* ("we, us") and *yu* ("you" singular) versus *yupela* ("you" plural). Another suffix, *-pela*, serves as a marker for adjectives of one syllable, demonstratives, indefinites, and numerals: *naispela* "pretty," *dispela* "this," *sampela* "some," *wanpela* "one." Verbs having a direct object (expressed or implied) take the suffix *-im*, and verbs without this suffix are intransitive or passive: *e.g.*, *rausim* "eject, remove" versus *raus* "be out, come out." Other parts of speech—nouns, prepositions, adverbs, conjunctions—are invariable but are distinguished by the types of combinations in which they occur. In other varieties of Pidgin English, the specific criteria for distinguishing classes of linguistic forms are different, but the basic structure is similar. Chinese Pidgin nouns and pronouns, for example, can take the locative suffix *-said*, in *doksaid* "at the dock (docks)," *maisaid* "at my house," and the verb suffix *-em* forms passive participles, as in *dis tri blong spoilem* "this tree is rotten."

Phrases and clauses

Syntax. The basic types of combination—phrases and clauses—found in pidgin are the same as those of English; here again, however, many details of syntax are different. Melanesian Pidgin nouns are followed not only by possessive phrases (*haus bilong mi* "house of me, my house") but also by modifying nouns (*haus pepa* "house [for] paper, office"), verbs (*haus kuk* "house [for] cooking, kitchen"), adverbs (*man nogud* "evil man," *nogud* being an adverb meaning "undesirably"), and clauses (*man mi lukim em* "the man I saw"). In Chinese Pidgin, pronouns simply precede nouns to indicate possession (*hi fes* "his face"), and relative position is shown by a noun preceding an adjective (*Ning-Po mo fa* "further than Ning Po"). With third person subjects, Melanesian Pidgin predicates are normally preceded by the predicate-marker *i-*: *ol i-singaut* "they call"; *balus i-no kamap yet* "the plane hasn't arrived yet." Melanesian Pidgin has the clause type called "equational" (also found in such languages as Russian and Hungarian) in which no verb is present: *dispela kaikai i-gudpela* "this food [is] good." Chinese Pidgin, on the other hand, has the copulative, or linking, verb *blong* "be" with nouns and adjectives in the predicate (*e.g.*, *yu fut blong plenti sor* "your foot is very sore") but uses no verb with an adverb in the predicate indicating location: *tumuchi dast tebal tapsaid* "a lot of dust [is] on the table."

Vocabulary

Lexicon. Since vocabulary is restricted (about 700 words in Chinese Pidgin, 2,000 in Melanesian), each word necessarily has a greater range of meaning than its English counterpart. The central meaning of Melanesian Pidgin *sori* is not "sorry" but "emotionally moved," as shown by its extension to "sympathetic, grateful, glad"; similarly, *dai* means "cease" ("die" is *dai tru* "stop for good"); and *stap* is "be located; remain; continue." For many concepts, pidgin uses phrases rather than single words: with *skru* "screw; joint" are formed *skru bilong arm* "elbow," *skru bilong leg* "knee," etc. Some pidgin words represent different parts of speech from their English counterparts; *e.g.*, the Melanesian Pidgin preposition *belong* "of" and the Chinese Pidgin copulative verb *blong* "be" from English "belong." Non-English meanings of pidgin words often reflect native social structure, as when *papa* means "uncle," since a boy's maternal uncle rather than his father (*papa tru*) is primarily responsible for his upbringing in New Guinea. Speakers of English are often naïvely amused or shocked by certain shifts of meaning, as when *ars* "buttocks" is extended to mean "bottom (of anything), foundation, reason, cause, source"; for example, *ars bilong diwai* "the base of the tree," or *God i-ars bilong olgeda samting* "God is the source of all things." In the context of native society and attitudes, however, these concepts are not taboo and no stigma attaches to the words or their use.

Non-English vocabulary. The proportion of vocabulary elements in Pidgin English derived from non-English sources is small. Of the approximately 2,000 words in Melanesian Pidgin, not over 10 percent are of non-English origin. Of these, perhaps half are Melanesian (such as *kiau* "eggbomb," *diwai* "tree," *malolo* "rest," *balus* "pigeon, airplane"), and one quarter were borrowed from German, such as *mark* "shilling," *tais* "pond," *langsam* "slow," and *beten* "prayer." The remainder are from various languages—a few from Malay (such as *karabau* "water buffalo") and three from Romance sources: *save* "know," *pikinini* "child," and *pato* "duck." The percentage of non-English elements in Chinese Pidgin is even smaller.

Structural changes from English

Restructuring. In the structural reduction from English to pidgin, the main grammatical characteristics have been kept (the part-of-speech system, the dichotomy between subject and predicate, the use of phrases functioning as single parts of speech), though often with different identifying features. The various kinds of Pidgin English are definitely English and Indo-European, not (as is often said) "native languages spoken with English words." A more sophisticated version of this latter theory is that indigenous vocabulary is simply replaced with new words by "relexification," with indigenous grammatical habits continuing. However, when new functional elements (*e.g.*, pronouns, inflectional suffixes, syntactic patterns) are also taken over, the process is one of complete language-substitution, involving replacement of grammatical structure ("regrammaticalization") as well. Nevertheless there have been extensive carryovers from non-English structural patterns because speakers of native languages have translated their own constructions into pidgin, especially in the early stages of its formation. For instance, in Chinese a numeral modifying a noun must be accompanied by a special word indicating a measure of quantity, called a "numeral-classifier," as in *sān ge rén*, literally "three piece man." This type of combination was reflected in older Chinese Pidgin *tufela man* "two men," *forpisi naif* "four knives," etc., with *-fela* for animate objects, *-pisi* for inanimate objects. However, in modern Chinese Pidgin, *-fela* has not survived, and *-pisi* has lost its independent status and become simply a numeral-suffix, as in *tupisi man* "two men." Similarly, the presence of separate pronouns for "we (excluding the hearer)" and "we (including the hearer)" in Melanesian languages has led to the establishment in Melanesian Pidgin of a parallel contrast between *mipela* "we (but not you)" and *yumi* "we (including you)."

Modern function. With the coming of modern civilization and technology to New Guinea and similar areas, pidgin has proved indispensable in education and political life. Earlier opposition to pidgin—partly on puristic, partly on anticolonialistic grounds—has proved unfounded. In New Guinea and the Solomons, and in many parts of West Africa, pidgin is no longer a status language or imposed on the people by white colonialists, but is the people's own lingua franca, indispensable for communication and easier to learn than English, which is both more complicated and more foreign to them. If skillfully used, pidgin can serve as both a medium of instruction and a bridge to English. In any case, it is clearly destined to remain as an increasingly useful lingua franca, with already manifest prospects of extensive creolization and resultant permanence as the native language of ever larger groups. (R.A.H.)

BIBLIOGRAPHY

General works. E.H. STURTEVANT, *Linguistic Change* (1917, reprinted 1961), was an important text in Indo-European comparative linguistics, until replaced by WINFRED P. LEHMANN, *Historical Linguistics: An Introduction* (1962). EDWARD SAPIR, *Language: An Introduction to the Study of Speech* (1921, reprinted 1957), presents a revision of the 19th-century whole-language typology. For an example of modern application of subsystem typology, see A.K. RAMANUJAN and COLIN MASICA, "Toward a Phonological Typology of the Indian Linguistic Area," in THOMAS A. SEBEOK (ed.), *Current Trends in Linguistics*, vol. 5 (1969), pp. 543–577; and for an example of lexical domain typology, see BRENT BERLIN and PAUL KAY, *Basic Color Terms: Their Universality and Evolution* (1969). ANTOINE MEILLET and MARCEL COHEN (eds.), *Les Langues du monde*, new ed. (1952), is a survey of languages of the world organized in terms of their genetic classification. Languages are also classified genetically in two more recent surveys: C.F. and F.M. VOEGELIN, "Languages of the World," in *Anthropological Linguistics*, vol. 6–8 (1964–66); and THOMAS A. SEBEOK (ed.), *Current Trends in Linguistics*, 13 vol. in 20 (1963–75)—individual volumes deal with geographic areas, and typological information on particular languages is presented. Variation in proposals for remote genetic relationships among the same set of languages is exemplified in two essays, both found in THOMAS A. SEBEOK (ed.), *Current Trends in Linguistics*, vol. 8, part 1 (1971): JOSEPH H. GREENBERG, "The Indo-Pacific Hypothesis," pp. 807–871; and STEPHEN A. WURM, "The Papuan Linguistic Situation," pp. 541–657. EDWARD SAPIR, "Central and North American Languages," in *Encyclopædia Britannica*, 14th ed. (1929), presented a classification of these languages in terms of six phyla; this classification was revised by a conference of specialists that resulted in the publication of C.F. and F.M. VOEGELIN (comps.), *Map of North American Indian Languages*, rev. ed. (1967). The culture areas of North American Indians are described in HAROLD E. DRIVER, *Indians of North America*, 2nd ed. rev. (1969). (C.F.V./F.M.V.)

Indo-European languages. KARL BRUGMANN, *Elements of the Comparative Grammar of the Indo-Germanic Languages*, 5 vol. (also published as *A Comparative Grammar of the Indo-Germanic Languages*, 1888–95; originally published in German, 1886–1900), is the latest completed full treatment of the whole family. A. MEILLET, *Introduction à l'étude comparative des langues indo-européennes*, 8th ed. corrected (1937, reissued 1978), is dated but still an excellent introduction to the subject. JERZY KURYŁOWICZ, *Indogermanische Grammatik*, ed. by MANFRED MAYRHOFER, especially vol. 1, part 1, *Einleitung*, by WARREN COWGILL (1986), and part 2, *Lautlehre (segmentale Phonologie des Indogermanischen)*, by MANFRED MAYRHOFER (1986), contains the most up-to-date account of the Indo-European sound system. ANDREW L. SIHLER, *New Comparative Grammar of Greek and Latin* (1995), focuses on the classical languages but provides the fullest introduction to Indo-European linguistics in English. JULIUS POKORNY, *Indogermanisches etymologisches Wörterbuch*, 2 vol. (1951–69), offers the most recent etymological dictionary of the whole family. CARL DARLING BUCK, *A Dictionary of Selected Synonyms in the Principal Indo-European Languages* (1949, reissued 1988), assembles a mine of information about Indo-European words for several hundred basic concepts. *The American Heritage Dictionary of Indo-European Roots*, rev. and ed. by CALVERT WATKINS (1985), focuses on the Indo-European component of English. HOLGER PEDERSEN, *Linguistic Science in the Nineteenth Century* (1931, reissued as *The Discovery of Language*, 1962; originally published in Danish, 1924), comprises a very good account of 19th-century work in the field. J.P. MALLORY, *In Search of the Indo-Europeans* (1989), provides a full and balanced account of the Indo-European homeland problem. The wide-ranging work by CALVERT WATKINS, *How to Kill a Dragon* (1995), studies aspects of Indo-European comparative poetics, religion, and mythology. (W.C./J.H.Ja.)

Anatolian languages. The best general handbook on Anatolian history and civilization, in German, is A. GOETZE, *Kleinasien*, 2nd ed. (1957), with succinct chapters on the Anatolian languages (pp. 45–63). An excellent, more popular account of Anatolia with marked emphasis on the Hittite period is O.R. GURNEY, *The Hittites*, rev. ed. (1961). This book deals with languages and races in ch. 6. The general handbook on the Anatolian languages, both Indo-European and non-Indo-European, written by scholars, is A. KAMMENHUBER *et al.*, *Altkleinasiatische Sprachen* (1969), although in some chapters the book is perhaps too detailed for the general reader. J. FRIEDRICH (ed.), *Kleinasiatische Sprachdenkmäler* (1932), is still a very valuable collection of texts, chosen to illustrate the various languages of Anatolia. Although written from the viewpoint of the Indo-Hittite hypothesis, E.H. STURTEVANT and E.A. HAHN, *A Comparative Grammar of the Hittite Language*, rev. ed. (1951), was a major achievement. In many respects it is still the best grammar of Hittite. H. PEDERSEN, *Hittitisch und die anderen Indoeuropäischen Sprachen* (1938), was directed against the Indo-Hittite hypothesis and should be used with Sturtevant and Hahn's book. J. PUHVEL's contribution "Dialectal Aspects of the Anatolian Branch of Indo-European," in H. BIRNBAUM and J. PUHVEL (eds.), *Ancient Indo-European Dialects*, pp. 235–247 (1966), gives a good survey of the main characteristics of the Anatolian subgroup. The best recent introduction to the "Indo-European problem" with a notable emphasis on Anatolian matters is R.A. CROSSLAND, *Immigrants from the North*, fasc. 60 of the *Cambridge Ancient History*, rev. ed. (1967). The best introduction to the Anatolian languages and to Hittite literature may be found in two highly authoritative and very well written scientific contributions by H.G. GUTERBOCK: "Toward a Definition of the Term Hittite," *Oriens*, 10:233–239 (1957) and "A View of Hittite Literature," *Journal of the American Oriental Society*, 84:107–115 (1964). The history of the various decipherments has been dealt with by J. FRIEDRICH, *Entzifferung verschollener Schriften und Sprachen*, 2nd ed. (1966). (Ph.H.J.H.t.C.)

Indo-Iranian languages. *Indo-Aryan languages (general works):* JULES BLOCH, *L'Indo-aryen du veda aux temps modernes* (1934; rev. Eng. trans., *Indo-Aryan from the Vedas to Modern Times*, 1965), a masterly survey of Indo-Aryan throughout its history; R.L. TURNER, *A Comparative Dictionary of the Indo-Aryan Languages* (1966), an indispensable source in which Sanskrit word headings are given Middle Indo-Aryan forms and New Indo-Aryan cognates; M.B. EMENEAU, "The Dialects of Old Indo-Aryan," in HENRIK BIRNBAUM and JAAN PUHVEL (eds.), *Ancient Indo-European Dialects*, pp. 123–138 (1966), a good summary, with discussion of proposed theories and references; SURYAKANTA, *A Practical Vedic Dictionary* (1981).

Old Indo-Aryan: THOMAS BURROW, *The Sanskrit Language*, new and rev. ed. (1973), a summary of the prehistory and history of Sanskrit, with references to Middle Indo-Aryan, which contains somewhat personal views but is valuable for its discussion of non-Aryan influences on Sanskrit; LOUIS RENOU, *Histoire de la langue sanskrite* (1956), an insightful summary of the grammar, vocabulary, and style of different stages of Sanskrit, with text selections and translations; MANFRED MAYRHOFER, *A Concise Etymological Sanskrit Dictionary*, 4 vol. (1953–80), contains sober etymologies, full references, and a discussion of loanwords and words supposed to have been borrowed from Dravidian.

Middle Indo-Aryan: RICHARD PISCHEL, *Grammatik der Prākrit-Sprachen* (1900; Eng. trans., *Comparative Grammar of the Prākrit Languages*, 2nd ed., 1965), an encyclopaedic grammar of all the Prākrits except Buddhist Hybrid Sanskrit and Pāli, which includes a good discussion of the different Prākrits in the introduction (now in need of updating); S.M. KATRE, *Prakrit Languages and Their Contribution to Indian Culture*, 2nd ed. (1964), a general survey of the Prākrits, including Pāli; LUDWIG ALSDORF, *Apabhraṃśa-Studien*, pp. 5–17, 20–37 (1937, reprinted 1966, important studies discussing noun and verb inflection.

Modern Indo-Aryan: JOHN BEAMES, *A Comparative Grammar of the Modern Aryan Languages: To Wit, Hindi, Panjabi, Sindhi, Gujarati, Marathi, Oṛiya, and Bangali*, 3 vol. (1872–79, reprinted 1966); and A.F.R. HOERNLE, *A Comparative Grammar of the Gaudian Languages with Special Reference to the Eastern Hindi* (1880, reprinted 1975), general comparative grammars of the New Indo-Aryan languages—though in need of modernization, still indispensable; GEORGE A. GRIERSON, *On the Modern Indo-Aryan Vernaculars* (1931), a reprint of two long articles, tracing the phonologic developments that led to New Indo-Aryan; S.K. CHATTERJEE, *Indo-Aryan and Hindi*, 2nd rev. ed.

(1960), a series of lectures briefly tracing the history of Indo-Aryan, with particular emphasis on Hindi and its relation to other Indo-Aryan languages and on the general language problem in India. (Ge.Ca.)

Iranian languages (general works): An important comprehensive treatment of the Iranian languages in general is the older study by WILHELM GEIGER and ERNST KUHN (eds.), *Grundriss der iranischen Philologie,* 2 vol. (1895–1904, reprinted 1974). This invaluable work is now in many respects antiquated, and it contains no account of several Middle Iranian languages that have been made known only in this century. A more recent account in less detail is provided by the *Handbuch der Orientalistik,* vol. 1, sect. 4, *Iranistik,* pt. 1, *Linguistik* (1958).

Old Iranian: A large number of Old Persian texts are given in transcription and translation in ROLAND KENT, *Old Persian,* 2nd ed. rev. (1953). Information on Old Persian linguistic problems is contained in WILHELM BRANDENSTEIN and MANFRED MAYRHOFER, *Handbuch des Altpersischen* (1964). On the Avestan language the article by GEORG MORGENSTIERNE, "Orthography and Sound-System of the Avesta," in *Norsk Tidsskrift for Sprogvidenskap,* 12:30–82 (1942), is of great importance.

Middle Iranian: The verbal system of Parthian is described by A. GHILAIN, *Essai sur la langue parthe, son système verbal d'après les textes manichéens* (1939, reprinted 1966); and for Middle Persian by W. HENNING, "Das Verbum des Mittelpersischen der Turfanfragmente," *Zeitschrift für Indologie und Iranistik,* 9:158–253 (1933). A Pahlavi dictionary for students is D.N. MACKENZIE, *A Concise Pahlavi Dictionary* (1971). The grammar of Sogdian has received detailed treatment in ILYA GERSHEVITCH, *A Grammar of Manichean Sogdian* (1954); and of Khotanese in R.E. EMMERICK, *Saka Grammatical Studies* (1968). A brief sketch of Khwārezmian is given in W.B. HENNING, "The Khwarezmian Language," *Zeki Velīdi Togan'a Armağan,* pp. 421–436 (1955).

Modern Iranian: A work of considerable importance is GILBERT LAZARD, *Grammaire du persan contemporain* (1957). For the early stages of Modern Persian, Lazard's *Langue des plus anciens monuments de la prose persane* (1963) is invaluable. The same author has provided a comprehensive guide to the most important linguistic features of Tadzhik in "Caractères distinctifs de la langue tadjik," *Bulletin de la Société linguistique de Paris,* 52:117–186 (1956). A treatment of Baluchi dialects is J.H. ELFENBEIN, *The Baluchi Language* (1966). Of the many works describing Kurdish dialects, a comprehensive modern work is D.N. MACKENZIE, *Kurdish Dialect Studies,* 2 vol. (1961–62). For the history of the Pashto language, GEORG MORGENSTIERNE, *An Etymological Vocabulary of Pashto* (1927), remains standard. Morgenstierne's *Indo-Iranian Frontier Languages,* 4 vol. in 6, 2nd rev. ed., with additional material (1973), is the work most frequently quoted for most of the minor languages. Ossetic has been described by a native speaker: V.I. ABAEV, *A Grammatical Sketch of Ossetic* (1964; originally published in Russian, 1959). (R.E.E.)

Greek language. LEONARD R. PALMER, *The Greek Language* (1980, reissued 1996), provides a historical overview.

Ancient Greek: Studies of Ancient Greek include MICHAEL VENTRIS and JOHN CHADWICK, *Documents in Mycenaean Greek,* 2nd ed. by JOHN CHADWICK (1973), covering both the writing system and the content of tablets from Knossos, Pylos, and Mycenae, by the authors of the decipherment; L.H. JEFFERY, *The Local Scripts of Archaic Greece,* rev. ed. by A.W. JOHNSTON (1990), a description of all the local varieties of the Greek alphabet, 8th–5th century BC; CARL DARLING BUCK, *The Greek Dialects* (1955, reprinted 1973), a summary of the dialectal features of Ancient Greek within the scope of a traditional descriptive grammar; A. MEILLET, *Aperçu d'une histoire de la langue grecque,* 8th ed. updated by OLIVIER MASSON (1975), the first and still fundamental endeavour to define the characteristics of Greek in a diachronic perspective; EDUARD SCHWYZER *et al., Griechische Grammatik,* 4 vol. in 5 (1934–71), and several later editions of various volumes, a complete description with exhaustive bibliography; PIERRE CHANTRAINE, *La Formation des noms en grec ancien* (1933, reissued 1979), dealing with the history of noun suffixes throughout the history of Greek, and *Dictionnaire étymologique de la langue grecque: histoire des mots,* 4 vol. in 5 (1968–80, reissued 4 vol. in 2, 1983–84), an excellent work including Mycenaean data; HJALMAR FRISK, *Griechisches etymologisches Wörterbuch,* 2nd ed., 3 vol. (1973–79), and a 3rd ed. (1991–), wisely selective (but often underrating Mycenaean data); and MICHEL LEJEUNE, *Phonétique historique du mycénien et du grec ancien* (1972), a thorough description of sound changes in Mycenaean and in the later Greek dialects. (M.Le./C.J.R.)

Koine and Byzantine Greek: A scholarly study of Koine Greek may be found in F. BLASS and A. DEBRUNNER, *A Greek Grammar of the New Testament and Other Early Christian Literature,* trans. and ed. by ROBERT W. FUNK (1961; originally published in German, 9th–10th ed., 1954–59), a classic work. FRANCIS THOMAS GIGNAC, *A Grammar of the Greek Papyri of the Roman and Byzantine Periods,* vol. 1, *Phonology* (1976), and vol. 2, *Morphology* (1981), provides a grammatical analysis of the documentary papyri and ostraca from Egypt (30 BC–AD 735) and offers comparison with other periods. ROBERT BROWNING, *Medieval and Modern Greek,* 2nd ed. (1983), covers the historical development of later periods, including a convenient summary of the development of Demotic.

Modern Greek: S.A. SOFRONIOU, *Teach Yourself Modern Greek* (1962, reissued as *Modern Greek,* 1993); and J.T. PRING, *A Grammar of Modern Greek on a Phonetic Basis* (1950, reissued 1975), are good elementary introductions. D.N. STAVROPOULOS, *Oxford Greek-English Learners Dictionary,* ed. by G.N. STAVROPOULOS (1988, reissued 1992), a dictionary of the spoken language, is accompanied by grammar tables. The structure and use of standard modern Greek are described in PETER MACKRIDGE, *The Modern Greek Language* (1985), which includes references to earlier works; and BRIAN D. JOSEPH and IRENE PHILIPPAKI-WARBURTON, *Modern Greek* (1987), which includes a discussion of certain theoretical issues. BRIAN NEWTON, *The Generative Interpretation of Dialect: A Study of Modern Greek Phonology* (1972), thoroughly describes modern Greek dialects in the framework of generative grammar. (B.E.N./A.M.-D.)

Italic languages. Detailed surveys of the Italic languages are found in J.H.W. PENNEY, "The Languages of Italy," in JOHN BOARDMAN *et al.* (eds.), *The Cambridge Ancient History,* vol. 4, 2nd ed. (1988), pp. 720–738; VITTORE PISANI, *Le lingue dell'Italia antica oltre il Latino,* 2nd ed., rev. and enlarged (1964); GIACOMO DEVOTO, *Gli antichi Italici,* 4th ed. (1969), and *The Languages of Italy* (1978; originally published in Italian, 1974); ALDO L. PROSDOCIMI (ed.), *Lingue e dialetti dell'Italia antica* (1978), with additions and indices published separately with the same title, ed. by ANNA MARINETTI (1982); and D. SILVESTRI, "Le lingue italiche," in ANNA GIACALONE RAMAT and PAOLO RAMAT (eds.), *Le lingue indoeuropee* (1993), pp. 349–371. The historical background to the Italic languages is treated by MASSIMO PALLOTTINO, *A History of Earliest Italy* (1991; originally published in Italian, 1984).

Fundamental works on Osco-Umbrian, most with selected texts, include ROBERT VON PLANTA, *Grammatik der oskisch-umbrischen Dialekte,* 2 vol. (1892–97, reissued 1973); CARL DARLING BUCK, *A Grammar of Oscan and Umbrian,* new ed. (1928, reprinted 1995); GINO BOTTIGLIONI, *Manuale dei dialetti italici: Osco, Umbro, e dialetti minori* (1954); A. ERNOUT, *Le dialecte ombrien: lexique du vocabulaire des "Tables Eugubines" et des inscriptions* (1961); and GERHARD MEISER, *Lautgeschichte der umbrischen Sprache* (1986). A detailed treatment of Paelignian, one of the central Oscan "minor dialects," is RAFAEL JIMÉNEZ ZAMUDIO, *Estudio del dialecto peligno y su entorno lingüístico* (1986). The most complete edition of Osco-Umbrian and Faliscan texts is EMIL VETTER, *Handbuch der italischen Dialekte,* vol. 1, *Texte mit Erklärung, Glossen, Wörterverzeichnis* (1953); it is supplemented by PAOLO POCCETTI, *Nuovi documenti italici* (1979). Special studies of some of the more important Oscan texts include ANNALISA FRANCHI DE BELLIS, *Le iovile capuane* (1981), and *Il cippo abellano* (1988). The most accurate text of the Iguvine Tables is found in ALDO L. PROSDOCIMI, *Le tavole iguvine,* vol. 1 (1984), with introduction and notes in Italian. Special studies of the Iguvine Tables include JAMES WILSON POULTNEY, *The Bronze Tables of Iguvium* (1959); and GIACOMO DEVOTO (ed.), *Tabulae Iguvinae,* rev. ed. (1940), in Latin. The authoritative edition of the South Picene texts is ANNA MARINETTI, *Le iscrizioni sudpicene* (1985–); IGNACIO-JAVIER ADIEGO LAJARA, *Protosabelio, osco-umbro, sudpiceno* (1992), in Spanish, is also useful. A grammatical description of Faliscan with many of the texts is provided by GABRIELLA GIACOMELLI, *La lingua falisca* (1963). Similar treatment is given to Venetic in G.B. PELLEGRINI and ALDO L. PROSDOCIMI, *La lingua venetica,* 2 vol. (1967); MICHEL LEJEUNE, *Manuel de la langue vénète* (1974); and GIULIA FOGOLARI *et al., I Veneti antichi: lingua e cultura* (1988). Important studies of onomastic material (so prominent in the Italic language texts) include, for Oscan, MICHEL LEJEUNE, *L'anthroponymie osque* (1976); and for Venetic, MICHEL LEJEUNE, *Ateste à l'heure de la romanisation (étude anthroponymique)* (1978); and JÜRGEN UNTERMANN, *Die venetischen Personennamen,* 2 vol. (1961). Questions of ancient Italic poetics are treated in CALVERT WATKINS, *How to Kill a Dragon: Aspects of Indo-European Poetics* (1995), especially chapters 10 and 17–20. (J.U./B.Vi.)

Romance languages. IORGU IORDAN and JOHN ORR, *An Introduction to Romance Linguistics,* 2nd ed. rev. with a supplement by REBECCA POSNER (1970; originally published in Romanian, 1962), describes 19th- and 20th-century scholarship and provides an extensive bibliography on the family as a whole and on individual languages; it may be supplemented by REBECCA POSNER and JOHN N. GREEN (eds.), *Trends in Romance Linguistics and Philology,* 5 vol. (1980–93). Useful works on the Romance languages in general include W.D. ELCOCK, *The Romance Languages,* new and rev. ed. revised by JOHN N. GREEN (1975), a philologically oriented work; REBECCA POSNER, *The Romance Languages* (1996), a linguistically oriented work; and MARTIN HARRIS and NIGEL VINCENT (eds.), *The Romance Languages* (1988). YAKOV MALKIEL, *Essays on Linguistic Themes* (1968), includes several articles on the Romance languages and Romance linguistics. On the individual languages, recommended works are L.R. PALMER, *The Latin Language* (1954, reprinted 1988); BRUNO MIGLIORINI, *The Italian Language,* abridged by T. GWYNFOR GRIFFITH (1966, reissued 1984; originally published in Italian, 1960); ALFRED EWERT, *The French Language,* 2nd ed. (1961); PETER RICKARD, *A History of the French Language,* 2nd ed. (1989, reissued 1993); R. ANTHONY LODGE, *French: From Dialect to Standard* (1993); WILLIAM J. ENTWISTLE, *The Spanish Language, Together with Portuguese, Catalan, and Basque,* 2nd ed. (1962); RALPH PENNY, *A History of the Spanish Language* (1991); and JOHN M. LIPSKI, *Latin American Spanish* (1994). (Re.P./M.S.)

Germanic languages. *History, classification, and description of the Germanic languages:* The developments from Indo-European to Proto-Germanic and on to the daughter languages are treated in HERMAN HIRT, *Handbuch des Urgermanischen,* 3 vol. (1931–34); EDUARD PROKOSCH, *A Comparative Germanic Grammar* (1939); HANS KRAHE, *Germanische Sprachwissenschaft,* 7th ed., edited by WOLFGANG MEID, 3 vol. (1967–69); LUDWIG ERICH SCHMITT (ed.), *Kurzer Grundriss der germanischen Philologie bis 1500,* 2 vol. (1970–71); FRANS VAN COETSEM and HERBERT L. KUFNER (eds.), *Toward a Grammar of Proto-Germanic* (1972); PAOLO RAMAT, *Introduzione alla linguistica germanica* (1980), also available in a German translation, *Einführung in das Germanische* (1981); JOSEPH B. VOYLES, *Early Germanic Grammar: Pre-, Proto-, and Post-Germanic Languages* (1992); and FRANS VAN COETSEM, *The Vocalism of the Germanic Parent Language: Systemic Evolution and Sociohistorical Context* (1994). The question of early dialect grouping is discussed in HANS FREDE NIELSEN, *The Germanic Languages: Origins and Early Dialectal Interrelations* (1989). Sketches of the earliest attested stages of the individual languages are given in ORRIN W. ROBINSON, *Old English and Its Closest Relatives: A Survey of the Earliest Germanic Languages* (1992); and discussions of earlier stages and descriptions of the modern languages are found in JOHAN VAN DER AUWERA and EKKEHARD KÖNIG (eds.), *The Germanic Languages* (1994).

Gothic: Grammars with reading selections are JOSEPH WRIGHT, *Grammar of the Gothic Language,* 2nd ed. (1954, reprinted 1981); WILLIAM H. BENNETT, *An Introduction to the Gothic Language* (1980); FERNAND MOSSÉ, *Manuel de la langue gotique,* new ed. adapted and expanded (1956); WOLFGANG KRAUSE, *Handbuch des Gotischen,* 3rd rev. ed. (1968); and WILHELM BRAUNE, *Gotische Grammatik,* 19th ed. rev. by ERNST A. EBBINGHAUS (1981). A discussion of Crimean Gothic is MACDONALD STEARNS, JR., *Crimean Gothic: Analysis and Etymology of the Corpus* (1978). A useful resource is WINFRED LEHMANN and HELEN-JO J. HEWITT, *A Gothic Etymological Dictionary* (1986).

Frisian: General works with discussions of historical and dialectal developments are THOMAS L. MARKEY, *Frisian* (1981); BO SJÖLIN, *Einführung in das Friesische* (1969); and PAOLO RAMAT, *Il Frisone: Introduzione allo studio della filologia frisone* (1967). Old Frisian grammars with reading selections are WILHELM HEUSER, *Altfriesisches Lesebuch mit Grammatik und Glossar* (1903, reprinted 1972); and WALTHER STELLER, *Abriss der altfriesischen Grammatik* (1928). Modern standard West Frisian is described in PIETER MEIJES TIERSMA, *Frisian Reference Grammar* (1985); J. ANGLADE, *Petit manuel de Frison moderne de l'ouest* (1966); and K. FOKKEMA, *Beknopte friese Spraakkunst,* 2nd ed. (1967).

Netherlandic: General discussions of the language are B.C. DONALDSON, *Dutch: A Linguistic History of Holland and Belgium* (1983); PIERRE BRACHIN, *The Dutch Language: A Survey* (1985; originally published in French, 1977); and C.G.N. DE VOOYS, *Geschiedenis van de Nederlandse Taal,* 5th rev. ed. (1952, reissued 1975). Historical grammars of the language are MORITZ SCHÖNFELD, *Schönfelds Historische Grammatica van het Nederlands,* 8th ed., edited by A. VAN LOEY (1970); JAN GOOSSENS, *Historische Phonologie des Niederländischen* (1974); C. VAN BREE, *Historische Grammatica van het Nederlands* (1987); and JOZEF VAN LOON, *Historische Fonologie van het Nederlands* (1986). Discussions of the earliest stage of Netherlandic are in ROLF H. BREMMER, JR., and AREND QUAK (eds.), *Zur Phonologie und Morphologie des Altniederländischen* (1992); and ANTHONY F. BUCCINI, "Onstaan en vroegste ontwikkeling van het Nederlandse taallandschap," *Taal en tongval,* 48 (Themanummer 8):8–66 (1995). Middle Dutch is treated in COLETTE M. VAN KERCKVOORDE, *An Introduction to Middle Dutch* (1993); JOHANNES FRANCK, *Mittelniederländische Grammatik,* 2nd rev. ed. (1910, reissued 1971); A. VAN LOEY, *Middelnederlandse spraakkunst,* 8th improved ed., 2 vol. (1976); and F.A. STOETT, *Middelnederlandsche spraakkunst: Syntaxis,* 3rd rev. ed. (1923). Grammars of the modern standard are B.C. DONALDSON, *Dutch Reference Grammar,* 3rd ed. (1987); M.C. VAN DEN TOORN, *Nederlandse grammatica,* 9th rev. ed. (1984); and G. GEERTS *et al.* (eds.), *Algemene Nederlandse spraakkunst* (1984).

Afrikaans: A comprehensive grammar of the standard language is B.C. DONALDSON, *A Grammar of Afrikaans* (1993). Treatments of the language's historical development are FRITZ PONELIS, *The Development of Afrikaans* (1993); and EDITH RAIDT, *Einführung in Geschichte und Struktur des Afrikaans* (1983).

German: The history of the language is covered in W.B. LOCKWOOD, *An Informal History of the German Language, with Chapters on Dutch and Afrikaans, Frisian, and Yiddish,* new ed. (1976); R.E. KELLER, *The German Language* (1978); JOHN T. WATERMAN, *A History of the German Language,* rev. ed. (1976, reissued 1991); C.J. WELLS, *German: A Linguistic History to 1945* (1985); and HUGO MOSER, *Deutsche Sprachgeschichte,* 6th rev. ed. (1969). Concise overviews of grammatical developments are CHARLES V.J. RUSS, *Historical German Phonology and Morphology* (1978), RICHARD VON KIENLE, *Historische Laut- und Formenlehre des Deutschen,* 2nd rev. ed. (1969); GÜNTHER SCHWEIKLE, *Germanische-deutsche Sprachgeschichte im Überblick,* 3rd improved and enlarged ed. (1990); and INGERID DAL, *Kurze deutsche Syntax auf historischer Grundlage,* 3rd improved ed. (1966). Treatments of older stages of the language are WILHELM BRAUNE, *Althochdeutsche Grammatik,* 14th ed., rev. by HANS EGGERS (1987); STEFAN SONDEREGGER, *Althochdeutsche Sprache und Literatur,* 2nd rev. and enlarged ed. (1987); and HERMANN PAUL, *Mittelhochdeutsche Grammatik,* 23rd ed., rev. by PETER WIEHL and SIEGFRIED GROSSE (1989). Grammars of the modern standard are GERHARD HELBIG and JOACHIM BUSCHA, *Deutsche Grammatik,* 16th ed. (1994); and GÜNTHER DROSDOWSKI *et al.* (eds.), *Der Duden in 10 Bänden,* 10 vol. in various editions (1984–91). Further useful works on the modern language are CHARLES V.J. RUSS, *The German Language Today: A Linguistic Introduction* (1994); and ANTHONY FOX, *The Structure of German* (1990). Descriptions of various dialects are found in CHARLES V.J. RUSS (ed.), *The Dialects of Modern German* (1989). Sociolinguistic aspects of German are discussed in STEPHEN BARBOUR and PATRICK STEVENSON, *Variation in German: A Critical Approach to German Sociolinguistics* (1990).

Old Saxon and Low German: A historical overview of linguistic developments in northern Germany is WILLY SANDERS, *Sachsensprache, Hansesprache, Plattdeutsch: sprachgeschichtliche Grundzüge des Niederdeutschen* (1982). Historical stages and dialects are described in JAN GOOSSENS (ed.), *Niederdeutsche: Sprache und Literatur,* vol. 1, *Sprache* (1973). Grammars of older stages are IRMENGARD RAUCH, *The Old Saxon Language* (1992); JOHAN HENDRIK GALLÉE, *Altsächsische Grammatik,* 3rd ed. (1993); and AGATHE LASCH, *Mittelniederdeutsche Grammatik,* 2nd ed. (1974). (Wi.G.M./A.F.Bu.)

Yiddish: URIEL WEINREICH, "Yiddish Language," in *Encyclopaedia Judaica,* vol. 16 (1971), columns 789–798; SOLOMON A. BIRNBAUM, *Yiddish: A Survey and a Grammar* (1979); and MAX WEINREICH, *History of the Yiddish Language* (1980; originally published in Yiddish, 1973), provide overviews. The main works up to 1958 are listed in URIEL WEINREICH and BEATRICE WEINREICH, *Yiddish Language and Folklore: A Selective Bibliography for Research* (1959). Significant works include *For Max Weinreich on His Seventieth Birthday: Studies in Jewish Languages, Literature, and Society* (1964); compilations of studies in language, folklore, and literature, all titled *The Field of Yiddish,* 1st collection ed. by URIEL WEINREICH (1954), 2nd collection, ed. by URIEL WEINREICH (1965), 3rd collection, ed. by MARVIN I. HERZOG, WITA RAVID, and URIEL WEINREICH (1969), 4th collection, ed. by MARVIN I. HERZOG *et al.* (1980), and 5th collection, ed. by DAVID GOLDBERG (1993); JOSHUA A. FISHMAN, *Yiddish in America: Socio-linguistic Description and Analysis* (1965); and MARVIN I. HERZOG, *The Yiddish Language in Northern Poland: Its Geography and History* (1965). URIEL WEINREICH, *Modern English-Yiddish, Yiddish-English Dictionary* (1968, reissued 1990); YUDA A. YOFE (JUDAH A. JOFFE) and YUDL MARK̦ (YUDEL MARK) (eds.), *Groyser verterbukh fun der Yidisher shprakh* (1961–); and MARVIN I. HERZOG (ed.), *The Language and Culture Atlas of Ashkenazic Jewry* (1992–), are useful reference works. (M.I.H.)

North Germanic (the Scandinavian languages): A survey of research on Scandinavian languages is presented in EINAR HAUGEN and THOMAS L. MARKEY, *The Scandinavian Languages: Fifty Years of Linguistic Research (1918–1968)* (1972). EINAR HAUGEN, *Scandinavian Language Structures* (1982), is a comparative survey of the languages. LARS S. VIKØR, *The Nordic Lan-*

guages: Their Status and Interrelations (1993), is a sociologically oriented introduction to the Scandinavian speech communities. Various important contributions, mainly on the grammar of the Scandinavian languages in the context of modern linguistics, are found in a series of conference volumes entitled *The Nordic Languages and Modern Linguistics* (irregular); the conferences began in 1969. ERNST HÅKON JAHR and OVE LORENTZ (eds.), *Phonology* (1981), *Prosody* (1983), *Morphology* (1985), *Syntax* (1989), and *Historical Linguistics* (1993), collects articles on the Norwegian language published during the last century and more. Compact histories of all the languages are presented in ELIAS WESSEN, *Die nordischen Sprachen,* trans. from Swedish (1968); and EINAR HAUGEN, *The Scandinavian Languages: An Introduction to Their History* (1976). Detailed histories include DIDRIK ARUP SEIP, *Norsk språkhistorie til omkring 1370,* 2nd ed. (1955), also available in a German translation, *Norwegische Sprachgeschichte,* ed. and expanded by LAURITS SALTVEIT (1971), for Old Norwegian; PETER SKAUTRUP, *Det danske sprogs historie,* 4 vol. and index (1944–70), for Danish; ELIAS WESSEN, *Svensk språkhistoria,* 3 vol. in various editions (1965), for Swedish; GUSTAV INDREBØ, *Norsk målsoga* (1951), for Norwegian; and EINAR HAUGEN, *Language Conflict and Language Planning* (1966), for the modern Norwegian period. JAN TERJE FAARLUND, *Syntactic Change: Toward a Theory of Historical Syntax* (1990), deals mainly with the Scandinavian languages in a theoretical perspective.

The classic grammars of the older languages are ADOLF G. NOREEN, *Altisländische und altnorwegische Grammatik,* 5th ed. (1970), and *Altschwedische Grammatik* (1904, reprinted 1978); and JOHANNES BRØNDUM-NIELSEN, *Gammeldansk grammatik i sproghistorisk fremstilling,* 5 vol. in various editions (1935–65). For introductory purposes the best grammar is SIGRID VALFELLS and JAMES E. CATHEY, *Old Icelandic: An Introductory Course* (1981). The most comprehensive grammars for the modern languages are PAUL DIDERICHSEN, *Elementær dansk grammatik,* 3rd ed. (1968), and his compendium *Essentials of Danish Grammar* (1964); and AAGE HANSEN, *Moderne dansk,* 3 vol. (1967), for Danish; ADOLF NOREEN, *Vårt språk,* 7 vol. (1903–24), for Swedish; JAN TERJE FAARLUND, SVEIN LIE, and KJELL IVAR VANNEBO, *Norsk referansegrammatikk* (1997), for Dano-Norwegian and New Norwegian; KRISTJÁN ÁRNASON, *Íslensk málfrædhi,* 2nd ed., 2 vol. (1982–83), for Icelandic; and W.B. LOCKWOOD, *An Introduction to Modern Faroese* (1955, reprinted 1977). Introductory textbooks for English-speaking users are ROBIN ALLAN, PHILIP HOLMES, and TOM LUNDSKAER-NIELSEN, *Danish: A Comprehensive Grammar* (1995); PHILIP HOLMES and IAN HINCHLIFFE, *Swedish: A Comprehensive Grammar* (1994); EINAR HAUGEN and KENNETH G. CHAPMAN, *Spoken Norwegian,* 3rd ed. (1982), for Dano-Norwegian; PETER HALLARÅKER, *Norwegian, Nynorsk: An Introduction for Foreign Students* (1983), for New Norwegian; and JÓN FREDHJÓNSSON, *A Course in Modern Icelandic,* 2 vol. (1978).

There are many-volumed native dictionaries for each language. Only some bilingual dictionaries are listed here: JENS AXELSEN (ed.), *The Standard Danish-English, English-Danish Dictionary* (1986); VINCENT PETTI *et al., Norstedts Comprehensive Swedish-English Dictionary,* 2nd ed. (1993), and *Norstedts Comprehensive English-Swedish Dictionary,* 2nd ed. (1994); WILLY A. KIRKEBY, *English-Norwegian Dictionary* (1995), and *Norsk-engelsk ordbok,* 2nd ed. (1986); EINAR HAUGEN (ed.), *Norwegian English Dictionary,* 4th ed. (1995), with entries in both the Norwegian languages; ARNGRÍMUR SIGURDHSSON, *Icelandic-English Dictionary,* 4th ed. (1983); G.T. ZOËGA, *English-Icelandic Dictionary,* 3rd ed. (1945, reprinted 1954); and ANNFINNUR Í SKÁLA, JONHARD MIKKELSEN, and ZAKARIAS WANG, *Ensk-føroysk ordhabók* (1992). (Ei.H./J.T.F.)

English language. *Dictionaries: The Oxford English Dictionary,* 2nd ed., 20 vol., ed. by JOHN A. SIMPSON and EDMUND S.C. WEINER (1989), incorporates all the words of the first edition and its supplementary volumes. Derivative dictionaries include *The Concise Oxford Dictionary of Current English,* 6th ed. (1976); *Oxford Illustrated Dictionary,* 2nd ed. (1975); *The Oxford Advanced Learner's Dictionary of Current English,* 3rd ed. (1974); and the *Oxford American Dictionary* (1980). Other one-volume dictionaries include *Chambers' Twentieth Century* (1972); *Longmans English Larousse* (1968); and P. HANKS, *Encyclopedic World Dictionary* (1971).

The leading American dictionary is *Webster's Third New International Dictionary of the English Language* (1961); it is updated by a separately published "Addenda" section, *6,000 Words* (1976). *Webster's New Collegiate Dictionary* (1979) is an abbreviated version. Other comprehensive dictionaries are *The New Century Dictionary of the English Language,* 2 vol. (1959); and *Funk and Wagnalls New Standard Dictionary of the English Language* (1963). Two comprehensive dictionaries are outstanding: *The Random House Dictionary of the English Language* (1966); and *The American Heritage Dictionary of the English Language* (1969).

Reliable etymological dictionaries include ERNEST WEEKLEY, *An Etymological Dictionary of Modern English,* 2 vol. (1921, reprinted 1967); E.H. PARTRIDGE, *Origins,* 5th rev. ed. (1971); and ERNEST KLEIN, *A Comprehensive Etymological Dictionary of the English Language,* 2 vol. (1966–67). *The Oxford Dictionary of English Etymology* (1966) will long remain the most authentic work in this field.

The two great historical dictionaries of American English are WILLIAM A. CRAIGIE and JAMES R. HULBERT (eds.), *A Dictionary of American English on Historical Principles,* 4 vol. (1936–44); and MITFORD M. MATHEWS, (ed.), *A Dictionary of Americanisms on Historical Principles,* 2 vol. (1951).

Modern usage: H.W. FOWLER, *A Dictionary of Modern English Usage* (1926), a somewhat eccentric work, has been thoroughly updated twice: 2nd ed. with the same title, rev. by ERNEST GOWERS (1965); and *The New Fowler's Modern English Usage,* 3rd ed. edited by R.W. BURCHFIELD (1996). It has its transatlantic counterpart in the following two works: BERGEN EVANS and CORNELIA EVANS, *A Dictionary of Contemporary American Usage* (1957); and MARGARET NICHOLSON, *A Dictionary of American-English Usage* (1957). See also ROY H. COPPERUD, *American Usage and Style* (1980).

Grammar and structure of English: A.A. HILL, *Introduction to Linguistic Structures: From Sound to Sentence in English* (1958); SAMUEL JAY KEYSER and PAUL M. POSRAL, *Beginning English Grammar* (1976); PAUL ROBERTS, *English Sentences* (1962); MARTIN JOOS, *The English Verb* (1964); H.A. GLEASON, *Linguistics and English Grammar* (1965); N.C. STAGEBERG, *An Introductory English Grammar,* 3rd ed. (1977); A.E. DARBYSHIRE, *A Description of English* (1967); R. QUIRK *et al., A Grammar of Contemporary English* (1972); B.M.H. STRANG, *Modern English Structure,* 2nd ed. rev. (1968); R.W. ZANDVOORT, *A Handbook of English Grammar,* 7th ed. (1975).

Phonetics of English: Handbooks include HANS KURATH and R.I. MCDAVID, *The Pronunciation of English in the Atlantic States* (1961); and A.C. GIMSON, *An Introduction to the Pronunciation of English* (1963).

Histories: An excellent "external history" is A.C. BAUGH, *A History of the English Language,* 3rd ed. (1978). FERNAND MOSSÉ, *Esquisse d'une histoire de la langue anglaise* (1947), is a masterpiece—brief, lucid, and profound. KARL BRUNNER, *Die englische Sprache: Ihre geschichtliche Entwicklung,* 2nd ed., 2 vol. (1960–62), is indispensable to advanced students.

Two brief surveys written early in the 20th century are recognized classics and remain stimulating: HENRY BRADLEY, *The Making of English* (1904, rev. by SIMEON POTTER, 1968); and J.O.H. JESPERSEN, *Growth and Structure of the English Language* (1905, reprinted 1971). Other fairly substantial histories include STUART ROBERTSON, *The Development of Modern English,* 2nd ed. rev. by FREDERIC G. CASSIDY (1954); M.M. BRYANT, *Modern English and Its Heritage,* 2nd ed. (1962); M.W. BLOOMFIELD and L.D. NEWMARK, *A Linguistic Introduction to the History of English* (1963); W.N. FRANCIS, *The English Language, an Introduction* (1965); THOMAS PYLES, *The Origins and Development of the English Language,* 2nd ed. (1971); SIMEON POTTER, *Our Language,* rev. ed. (1968); J.W. CLARK, *Early English: A Study of Old and Middle English* (1967); A.C. PARTRIDGE, *Tudor to Augustan English* (1969); J.A. SHEARD, *The Words We Use,* rev. ed. (1970); JOSEPH M. WILLIAMS, *Origins of the English Language: A Social and Linguistic History* (1975); and B.M.H. STRANG, *A History of English* (1970). F.T. VISSER, *An Historical Syntax of the English Language,* 3 vol. (1963–73), provides copious illustrations and bibliographies.

Special studies: GEORGE W. TURNER, *The English Language in Australia and New Zealand* (1966); SIMEON POTTER, *Changing English* (1969); JOHN W. SPENCER (ed.), *The English Language in West Africa* (1971); MITFORD M. MATHEWS (ed.), *The Beginnings of American English* (1931); THOMAS PYLES, *Words and Ways of American English* (1952); ALBERT H. MARCKWARDT, *American English,* 2nd. ed. rev. by J.L. DILLARD (1980); and *Black English: Its History and Usage in the United States* (1972), also by Dillard. (S.P.)

Armenian language. S.L. KOGIAN, *Armenian Grammar* (1949); H. HUBSCHMANN, "Armenische Grammatik," *Armenische Etymologie* (1897); A. MEILLET, *Esquisse d'une grammaire comparée de l'arménien classique,* new ed. (1936); GERHARD DEETERS, *Armenisch und Südkaukasisch* in *Caucasica* (1927); HEINRICH ZELLER, "Armenisch," in *Geschichte der indogermanischen Sprachwissenschaft,* vol. 4 (1927). For ancient Armenian, see A. MEILLET, *Altarmenisches Elementarbuch* (1913); and H. JENSEN *Altarmenische Grammatik* (1959). Medieval Armenian is treated in J. KARST, *Historische Grammatik des Kilikisch-Armenischen* (1901). For modern speech, see H. ADJARIAN, *Classification des dialectes arméniens* (1909); and A. ABEGHIAN, *Neuarmenische Grammatik* (1936). (H.K.V.)

Tocharian languages. DOUGLAS Q. ADAMS, *Tocharian Historical Phonology and Morphology* (1988), is an introduction to

Tocharian from the Indo-European point of view; though HOLGER PEDERSEN, *Tocharisch vom Gesichtspunkt der indoeuropäischen Sprachvergleichung* (1941), is also still of value. EMIL SIEG and WILHELM SIEGLING (eds.), *Tocharische Sprachreste,* 1 vol. in 2 (1921), gives the transcription of all the manuscripts in Tocharian A and reproduces a number of the best-preserved leaves in facsimile. WILHELM SCHULZE, EMIL SIEG, and WILHELM SIEGLING, *Tocharische Grammatik* (1931), is an exhaustive grammar of Tocharian A, with a verbal index identifying and listing all verb forms in that language. WOLFGANG KRAUSE, *Westtocharische Grammatik,* vol. 1, *Das Verbum* (1952), is indispensable for the verb in Tocharian B. WOLFGANG KRAUSE and WERNER THOMAS, *Tocharisches Elementarbuch,* 2 vol. (1960–64), comprises a grammar, glossary, and texts for both languages; while A.J. VAN WINDEKENS, *Le tokharien confronté avec les autres langues indo-européennes,* 2 vol. (1976–82), provides a general, if at times idiosyncratic, etymological dictionary and comparative phonology and morphology of both languages. GEORGE S. LANE, "On the Interrelationship of the Tocharian Dialects," in HENRIK BIRNBAUM and JAAN PUHVEL (eds.), *Ancient Indo-European Dialects* (1966), pp. 213–233, attempts to solve some of the problems concerning the varied uses of the two languages. (G.S.L./D.Q.A.)

Celtic languages. H. LEWIS and H. PEDERSEN, *A Concise Comparative Celtic Grammar* (1937), is a survey of the entire field. The early history of the British group is discussed in detail by K.H. JACKSON, *Language and History in Early Britain* (1953); for the later history of Welsh and Breton, J. MORRIS JONES, *A Welsh Grammar* (1913), and F. GOURVIL, *Langue et littérature bretonnes* (1952), give useful information. P. BERRESFORD ELLIS, *The Cornish Language and Its Literature* (1974), deals with the early period, as well as with the recent Cornish language revival movement. R. THURNEYSEN, *Grammar of Old Irish,* rev. ed. (1946), is a classic among linguistic handbooks; for the later development of Irish, T.F. O'RAHILLY, *Irish Dialects Past and Present* (1932), is full of information and contains chapters on Scottish Gaelic and Manx. B.O CUIV (ed.), *A View of the Irish Language* (1969), containing 12 essays by various hands, maps, and illustrations; and D. GREENE, *The Irish Language* (1966), are directed to the general reader rather than to the linguist. The relevant sections of GLANVILLE PRICE, *The Present Position of Minority Languages in Western Europe* (1969), give full bibliographies of works dealing with the political and social status of the surviving Celtic languages. (D.Gr.)

Baltic languages. There are very few works on the Baltic languages in English aside from LEONARDAS DAMBRIŪNAS, ANTANAS KLIMAS, and WILLIAM R. SCHMALSTIEG, *Introduction to Modern Lithuanian* (1966); TERĒZA BUDIŅA LAZDIŅA, *Teach Yourself Latvian* (1966); and JANIS ENDZELĪNS, *Baltų kalbų garsai ir formos* (1957; Eng. trans., *Comparative Phonology and Morphology of the Baltic Languages,* 1971). *Baltic Linguistics,* ed. by THOMAS F. MAGNER and WILLIAM R. SCHMALSTIEG (1970), is a collection of papers on various aspects of Baltic linguistics.

Works in other languages include REINHOLD TRAUTMANN, *Die altpreussischen Sprachdenkmäler* (1910); JOHANN ENDZELIN, *Lettische Grammatik* (1922; trans. into Latvian, 1951); KAZIMIERAS BŪGA, *Lietuvių kalbos žodynas* (1924–25), the introduction to this dictionary containing much valuable information on the history of the Baltic languages; *Rinktiniai raštai,* 3 vol. (1958–61); JANIS ENDZELĪNS, *Senprūšu valoda* (1943; trans. into German, 1944, without a glossary); *Ievads baltu filoloģijā* (1945), an introduction to Baltic linguistics, in Latvian; ERNST FRAENKEL, *Die baltischen Sprachen* (1950), a general introduction; ALFRED SENN, "Die Beziehungen des Baltischen zum Slavischen und Germanischen," *Zeitschrift für vergleichende Sprachforschung,* vol. 71 (1954), discusses the relationship of Baltic to Slavic and Germanic; *Mūsdienu latviešu literārās valodas gramatika,* 2 vol. (1959–62), a grammar of the modern Latvian literary language; ARTURS OZOLS, *Veclatviešu rakstu valoda* (1965), treats Old Latvian; JAN OTREBSKI, *Gramatyka języka litewskiego,* 3 vol. (1956–65), a Lithuanian grammar, in Polish; MARTA RUDZĪTE, *Latviešu dialektoloģija* (1964), treats Latvian dialectology; *Lietuvių kalbos gramatika,* 2 vol. (1956, 1971), the most authoritative grammar of the Lithuanian language; ZIGMAS ZINKEVICIUS, *Lietuvių dialektologija* (1966), a valuable treatment of Lithuanian dialectology; CHRISTIAN S. STANG, *Vergleichende Grammatik der Baltischen Sprachen* (1966), the only scholarly comparative grammar of the Baltic languages; VYTAUTAS J. MAZIULIS (comp.), *Prūsų kalbos paminklai* (1966), contains and discusses all the photographed Old Prussian texts; ALGIRDAS SABALIAUSKAS, "Lietuvių kalbos leksikos raida," *Lietuvių kalbotyros klausimai,* 8:5–140 (1966), treats the development of the vocabulary of Lithuanian; JONAS PALIONIS, *Lietuvių literatūrinė kalba XVI–XVII a.* (1967), a treatment of the Lithuanian literary language in the 16th and 17th centuries; JONAS KAZLAUSKAS, *Lietuvių kalbos istorinė gramatika* (1968), the only historical grammar of Lithuanian; VYTAUTAS MAZIULIS, *Baltų ir kitų indoeuropiečių kalbų santykiai* (1970), treats the relationship of Baltic and the other Indo-European languages.

The largest dictionaries of Baltic languages are K. MULENBACHS, *Latviešu valodas vārdnīca,* 4 vol. (1923–32); *Lietuvių kalbos žodynas,* 8 vol. (1941–70); and ERNST FRAENKEL, *Litauisches etymologisches Wörterbuch,* 2 vol. (1955–65). (V.J.M.)

Slavic languages. BERNARD COMRIE and GREVILLE G. CORBETT (eds.), *The Slavonic Languages* (1993), describes each language according to a set plan, adding historical sections. An older reference is R.G.A. DE BRAY, *Guide to the Slavonic Languages,* 3rd ed., rev. and expanded, 3 vol. (1980). ROMAN JAKOBSON, *Slavic Languages,* 2nd ed. (1955, reprinted 1963), is a masterful, brief structural sketch. ALEXANDER M. SCHENKER, *The Dawn of Slavic* (1996), combines introductions to the early literature and history of the Slavs (with extracts from source texts) and a Proto-Slavic grammar. The classic introduction to Proto-Slavic (common Slavic) from the viewpoint of Indo-European is A. MEILLET, *Le Slave commun,* 2nd ed. rev. and enlarged with A. VAILLANT (1934, reprinted 1965). GEORGE Y. SHEVELOV, *A Prehistory of Slavic* (1964), treats phonology in detail. CHARLES E. TOWNSEND and LAURA A. JANDA, *Common and Comparative Slavic: Phonology and Inflection* (1996); and TERENCE R. CARLTON, *Introduction to the Phonological History of the Slavic Languages* (1991), are textbooks that also trace post-Common Slavic developments. HENRIK BIRNBAUM, *Common Slavic: Progress and Problems in Its Reconstruction* (1975); and HENRIK BIRNBAUM and PETER T. MERRILL, *Recent Advances in the Reconstruction of Common Slavic (1971–1982)* (1985), list and comment on the extensive scholarly literature. KAREL HORÁLEK, *An Introduction to the Study of the Slavonic Languages,* trans. and amended by PETER HERRITY, 2 vol. (1992; originally published in Czech, 2nd enlarged ed., 1962), treats many topics including the rise of standard languages; this latter subject is particularly addressed in ALEXANDER M. SCHENKER, EDWARD STANKIEWICZ, and MICAELA S. IOVINE (eds.), *The Slavic Literary Languages* (1980). Hypotheses on the homeland and migrations of the pre-Slavs are discussed in the light of linguistic evidence—*i.e.,* vocabulary and names—in ZBIGNIEW GOŁĄB, *The Origins of the Slavs: A Linguist's View* (1992); while MARIJA GIMBUTAS, *The Slavs* (1971), provides an archaeological survey and discusses religious and social vocabulary. There is no complete Slavic etymological dictionary, though several have been begun—*e.g.,* O.N. TRUBACHEV (ed.), *Etimologicheskiĭ slovar' slavianskikh iazykov* (1974–); other useful dictionaries are MAX VASMER, *Russisches etymologisches Wörterbuch,* 3 vol. (1950–58); or, for Old Church Slavonic vocabulary, LINDA SADNIK and RUDOLF AITZETMÜLLER, *Handwörterbuch zu den altkirchenslavischen Texten* (1955, reissued 1989). FRANTIŠEK KOPEČNÝ, *Základní všeslovanská slovní zásoba* (1981), lists 2,000 words that occur in every Slavic language. Slavic material has been important in the development of general linguistics; MORRIS HALLE (ed.), *Roman Jakobson: What He Taught Us* (1983), illustrates structural approaches; and STEVEN FRANKS, *Parameters of Slavic Morphosyntax* (1994), focuses on generative methods. (V.V.I./W.Br.)

Albanian language. P.L. HORECKY (ed.), *Southeastern Europe: A Guide to Basic Publications,* pp. 102–111 (1969), including a compilation by E.P. HAMP of about 50 major annotated items on Albanian scholarship, language, literature, folklore, ethnography, and folk music, with references to other supporting work; E.P. HAMP, "Albanian," in THOMAS A. SEBEOK (ed.), *Current Trends in Linguistics,* vol. 9 (1972), pp. 1626–92, a review of work since 1918, with copious bibliography; LEONARD NEWMARK, P. HUBBARD, and PETER PRIFTI, *Standard Albanian* (1982), a reference grammar. (E.P.H.)

Uralic languages. The following manuals provide a sound basis for further study: BJÖRN COLLINDER, *Fenno-Ugric Vocabulary,* 2nd, rev. ed. (1977), presenting comparative Uralic word lists; BJÖRN COLLINDER *et al.* (compilers), *Survey of the Uralic Languages,* 2nd, rev. ed. (1969), which gives short sketches of all but a few of the Uralic languages, with the lesser languages receiving only superficial treatment; DENIS SINOR (ed.), *The Uralic Languages* (1988), a collection of articles (many in English) on Uralic topics; TOIVO VUORELA, *The Finno-Ugric Peoples,* trans. from Finnish (1964), an anthropological survey; PETER HAJDU, *Finno-Ugrian Languages and Peoples* (1975; originally published in Hungarian, 1962); and MORRIS SWADESH, "Linguistic Overview," in JESSE D. JENNINGS and EDWARD NORBECK (eds.), *Prehistoric Man in the New World* (1964, reissued 1974), pp. 527–556, which posits Uralic–American Indian relationships.

Works dealing with specific languages include PETER HAJDU, *The Samoyed Peoples and Languages,* trans. from Hungarian, 2nd rev. ed. (1968); ROBERT T. HARMS, *Estonian Grammar* (1962); THOMAS A. SEBEOK and FRANCES J. INGEMANN, *An Eastern Cheremis Manual* (1961), a clear, concise description of one of the lesser languages; FRED KARLSSON, *Finnish Gram-*

mar, 2nd ed. (1987; originally published in Swedish, 1978); and ZOLTÁN BÁNHIDI, ZOLTÁN JÓKAY, and DÉNES SZABÓ, *Learn Hungarian,* 6th ed. (1988). (R.T.H.)

Altaic languages. N. POPPE, *Introduction to Altaic Linguistics* (1965), outlines the history and structure of the languages (including Korean), as well as the history of scholarship in Altaic linguistics, while his *Vergleichende Grammatik der altaischen Sprachen,* vol. 1, *Vergleichende Lautlehre* (1960), is a comprehensive comparative phonology. G.J. RAMSTEDT, *Einführung in die altaische Sprachwissenschaft,* ed. by PENTTI AALTO, 3 vol. (1952–66), is a classic of comparative grammar, although it is somewhat outdated. *Handbuch der Orientalistik,* vol. 5 in 3 parts (1963–68), collects essays on languages in the Turkic, Mongolian, and Manchu-Tungus families. The locations of the various languages are shown in two map supplements to *National Geographic,* "Peoples of the Soviet Union" (February 1976), and "The Peoples of China" (July 1980). The classic though now out-of-date work on the relationship of Korean and Altaic is G.J. RAMSTEDT, *Studies in Korean Etymology,* 2 vol. (1949–53). The relationship of Japanese and Altaic is proposed and argued for in ROY ANDREW MILLER, *Japanese and the Other Altaic Languages* (1971).

A good, if uneven, handbook on the Turkic family is KARL H. MENGES, *The Turkic Languages and Peoples,* 2nd rev. ed. (1995). Descriptions of the individual languages are contained in *Philologiae Turcicae Fundamenta,* vol. 1 (1959). N. POPPE, *Introduction to Mongolian Comparative Studies* (1955), includes a history and description of the various languages and a brief history of scholarship. JOHANNES BENZING, *Die tungusischen Sprachen* (1956), is a classic work on Manchu-Tungus languages. (G.Ha./Ed.)

Dravidian languages. R. CALDWELL, *A Comparative Grammar of the Dravidian or South-Indian Family of Languages,* 3rd ed. by J.L. WYATT and T.R. PILLAI (1913, reprinted 1956 and 1961), the classic work that laid the foundations of Dravidian linguistics; G.A. GRIERSON (ed.), *Linguistic Survey of India,* vol. 4, *Muṇḍā and Dravidian Languages,* by S. KONOW (1906); T. BURROW and M.B. EMENEAU, *A Dravidian Etymological Dictionary* (1961, reprinted 1966; *Supplement,* 1968), the first etymological dictionary of the family, marking a new era in Dravidian studies (indispensable point of departure for any further work in the field); B. KRISHNAMURTI, *Telugu Verbal Bases: A Comparative and Descriptive Study* (1961), an indispensable study of the phonology and derivational morphology of Dravidian, and "Comparative Dravidian Studies," in THOMAS A. SEBEOK (ed.), *Current Trends in Linguistics,* vol. 5 (1969), pp. 309–333, a summary treatment of the latest developments in the field; K. ZVELEBIL, *Comparative Dravidian Phonology* (1970), the first systematic compendium of the comparative phonology of Dravidian; J. BLOCH, *The Grammatical Structure of Dravidian Languages* (1954; originally published in French, 1946), an excellent description of the main morphological and syntactic features of the family that ignores phonology totally; F.B.J. KUIPER, "The Genesis of a Linguistic Area," *Indo-Iranian Journal,* 10:81–102 (1967), a treatment of the problems of Aryan and Dravidian convergence. (K.V.Z.)

Austroasiatic languages. H.L. SHORTO, JUDITH M. JACOB, and E.H.S. SIMMONDS (compilers), *Bibliographies of Mon-Khmer and Tai Linguistics* (1963), is an unannotated bibliography of linguistic books and articles from 1790 to 1960 that does not include the Munda subfamily or the Viet-Muong branch but incorporates the (Austronesian) Cham languages into Mon-Khmer. Two works by W. SCHMIDT, *Grundzüge einer Lautlehre der Mon-Khmer-Sprachen* (1906), and *Die Mon-Khmer-Völker: ein Bindeglied zwischen Völkern Zentralasiens und Austronesiens* (1906), also available in a French translation, "Les Peuples Mon-Khmer: trait-d'union entre les peuples de l'Asie Centrale et de l'Austronésie," *Bulletin de l'École française d'Extrême-Orient,* 7:213–263 and 8:1–35 (1907–08), for the first time support the Austroasiatic hypothesis with lexical, phonological, and morphological evidence; they remain the basic work of Austroasiatic studies. HEINZ-JÜRGEN PINNOW, *Versuch einer historischen Lautlehre der Kharia-Sprache* (1959), an ambitious project with somewhat uncertain results, contains an analysis and systematic comparison of the phonologies of Munda languages and establishes connections with the rest of the Austroasiatic group. *Mon-Khmer Studies* (irregular) collects short technical articles mostly on the Montagnard languages of southern Vietnam, with topics varying from basic vocabulary to phonology, morphology, syntax, folk taxonomies, and oral literature. (G.Di.)

Sino-Tibetan languages. *General works:* Overviews are provided in PAUL K. BENEDICT, *Sino-Tibetan: A Conspectus* (1972); G.A. GRIERSON (compiler and ed.), *Linguistic Survey of India,* vol. 3 in 3 parts, *Tibeto-Burman Family* (1903–09, reprinted 1967); FRANK M. LEBAR, GERALD C. HICKEY, and JOHN K. MUSGRAVE, *Ethnic Groups of Mainland Southeast Asia* (1964); HENRI MASPERO, "Langues de l'Asie du Sud-Est," in A. MEILLET and MARCEL COHEN (eds.), *Les Langues du monde,* new ed. (1952, reprinted 1981), pp. 525–644, the most authoritative concise treatment of Sino-Tibetan; C.F. VOEGELIN and F.M. VOEGELIN, "Languages of the World: Sino-Tibetan," *Anthropological Linguistics,* vol. 6, no. 3 (March 1964) and vol. 7, nos. 3–6, all part 1 (March–June 1965); ROBERT SHAFER, *Introduction to Sino-Tibetan,* 5 vol. (1966–74); ROBERT SHAFER (ed.), *Bibliography of Sino-Tibetan Languages,* 2 vol. (1957–63), indispensable for further research; and S. ROBERT RAMSEY, *The Languages of China* (1987), comprehensive and not difficult to read. WELDON SOUTH COBLIN, *A Sinologist's Handlist of Sino-Tibetan Lexical Comparisons* (1986), is a competent statement on comparative work; to some extent it builds on NICHOLAS C. BODMAN, "Proto-Chinese and Sino-Tibetan," in FRANS VAN COETSEM and LINDA R. WAUGH, *Contributions to Historical Linguistics* (1980), pp. 34–199.

Chinese: Three pertinent essays are found in THOMAS A. SEBEOK (ed.), *Current Trends in Linguistics,* vol. 2 (1967): NICHOLAS C. BODMAN, "Historical Linguistics," pp. 3–58; KUN CHANG, "Descriptive Linguistics," pp. 59–90; and the technical work by SØREN EGEROD, "Dialectology," pp. 91–129. Grammars include YUEN REN CHAO, *Mandarin Primer* (1948, reissued 1976), excellent chapters on script and grammar; YUEN REN CHAO and LIEN SHENG YANG, *Concise Dictionary of Spoken Chinese* (1947, reissued 1970), with an excellent introduction; YUEN REN CHAO, *A Grammar of Spoken Chinese* (1968), a reliable standard reference work; W.A.C.H. DOBSON, *Late Archaic Chinese* (1959), and *Early Archaic Chinese* (1962), useful but not always trustworthy guides to Old Chinese grammar; EDWIN G. PULLEYBLANK, *Outline of Classical Chinese Grammar* (1995), a concise guide to usage in the texts; and CHRISTOPH HARBSMEIER, *Aspects of Classical Chinese Syntax* (1981), an important, pioneering contribution. Language reform is addressed in JOHN DEFRANCIS, *Nationalism and Language Reform in China* (1950, reissued 1972), informative and readable, and "Language and Script Reform," in THOMAS A. SEBEOK (ed.), *Current Trends in Linguistics,* vol. 2 (1967), pp. 130–150. R.A.D. FORREST, *The Chinese Language,* 3rd ed. (1973), a standard reference work, also treats related and contiguous languages. A variety of topics is covered in these works of BERNHARD KARLGREN: *Études sur la phonologie chinoise* (1915–26), an epoch-making work but very technical, *Sound and Symbol in Chinese,* rev. ed. (1962, reissued 1971), very readable, but somewhat out of date, *The Chinese Language: An Essay on Its Nature and History,* trans. from Swedish (1949), a popular account of phonetic reconstructions, and *Easy Lessons in Chinese Writing* (1958), an interesting account of the etymology of Chinese characters. PAUL KRATOCHVIL, *The Chinese Language Today: Features of an Emerging Standard* (1968), is very readable. A more detailed description of the language is found in JERRY NORMAN, *Chinese* (1988). MANTARO J. HASHIMOTO, *Phonology of Ancient Chinese,* 2 vol. (1978–79), is rather technical. EDWIN G. PULLEYBLANK, *Middle Chinese: A Study in Historical Phonology* (1984), is very important but not easily read, and information on the earlier pronunciations of specific characters can be found in his *Lexicon of Reconstructed Pronunciation in Early Middle Chinese, Late Middle Chinese, and Early Mandarin* (1991). An important and very readable statement on Old Chinese is FANG KUEI LI, "Archaic Chinese," in DAVID N. KEIGHTLEY (ed.), *The Origins of Chinese Civilization* (1983), pp. 393–408.

Tibeto-Burman and Karen: JAMES A. MATISOFF, "The Languages and Dialects of Tibeto-Burman," in JOHN MCCOY and TIMOTHY LIGHT (eds.), *Contributions to Sino-Tibetan Studies* (1986), pp. 3–75, lists and comments on all known Tibeto-Burman idioms. Two informative overview essays are found in THOMAS A. SEBEOK (ed.), *Current Trends in Linguistics:* KUN CHANG, "National Languages," vol. 2 (1967), pp. 151–176, treating minority languages in China; and ROY ANDREW MILLER, "The Tibeto-Burman Languages of South Asia," vol. 5 (1969), pp. 431–449. STUART N. WOLFENDEN, *Outlines of Tibeto-Burman Linguistic Morphology* (1929), is the classic statement. Works on specific languages include JAMES A. MATISOFF, *The Grammar of Lahu,* updated ed. (1982), the best grammar of a Lolo-type language; ROBERT B. JONES, JR., and U KHIN, *The Burmese Writing System* (1953); ROBERT B. JONES, JR., *Karen Linguistic Studies* (1961), a technical treatment; and ROY ANDREW MILLER, *The Tibetan System of Writing* (1956), a brief study. (S.C.E.)

Tai languages. Studies include MARY R. HAAS, *Thai-English Student's Dictionary* (1964), a concise dictionary arranged according to the order of the Thai alphabet with a brief description of the phonological and grammatical system; MARY R. HAAS and HENG R. SUBHANKA, *Spoken Thai,* 2 vol. (1945, reissued 1978), a textbook; GEORGE BRADLEY MCFARLAND, *Thai-English Dictionary* (1941, reprinted 1974), with more extensive lexical coverage but with a nonstandard transcription of Thai pronunciation; FANG KUEI LI (FANG-KUEI LI), *A Handbook of Comparative Tai* (1977), which lists more than 1,000 sets of cognate words in three Tai languages and gives reconstructions of the initial consonants, vowels, and tones of the ancestral Proto-Tai language, and "The Tai and the Kam-Sui Languages," *Lingua,* 14:148–179 (1965), a discussion of the relationship of these two language groups; RICHARD B. NOSS, *Thai Reference Grammar* (1964), a detailed descriptive and structural analysis; PAUL K. BENEDICT, *Austro-Thai Language and Culture, with a Glossary of Roots* (1975), an attempt to show a relationship between the Tai(-Kadai) languages and the Austronesian languages; and TATSUO HOSHINO and RUSSELL MARCUS, *Lao for Beginners: An Introduction to the Spoken and Written Language of Laos* (1981). Lexical resources on other Tai languages such as Yay, Lue, Saek, and regional dialects include the various volumes in the *Michigan Papers on South and Southeast Asia* series by WILLIAM J. GEDNEY, ed. by THOMAS JOHN HUDAK, each of which contains an extensive glossary and transcriptions of either texts or example sentences. (F.K.L./D.B.So.)

Paleo-Siberian languages. DEAN WORTH, "Paleosiberian," in THOMAS A. SEBEOK (ed.), *Current Trends in Linguistics,* vol. 1 (1963), pp. 345–373, provides good coverage of Soviet work after World War II. ROMAN JAKOBSON, GERTA HÜTTL-WORTH, and JOHN FRED BEEBE, *Paleosiberian Peoples and Languages: A Bibliographical Guide* (1957, reprinted 1981), is very useful, with an informative appendix. Still a good source on Paleo-Siberian languages is V.V. VINOGRADOV (ed.), *IAzyki narodov SSSR,* vol. 5, *Mongol'skie, tunguso-man'chzhurskie i paleoaziatskie iazyki,* ed. by P.IA. SKORIK *et al.* (1968). More recent studies, also in Russian, are the articles in N.I. KONRAD *et al.* (eds.), *IAzyki Azii i Afriki,* vol. 3, *IAzyki drevnei Perednei Azii (nesemitskie)* (1979), in which Nivkh and Ket in particular are presented in more detail; a reconstruction of proto-Yeniseic by S.A. STAROSTIN, "Praeniseiskaia rekonstruktsiia i vneshnie sviazi eniseĭskikh iazykov," in V.V. IVANOV, V.N. TOPOROV, and B.A. USPENSKII (eds.), *Ketskii sbornik,* vol. 3, *Antropologiia, etnografiia, mifologiia, lingvistika* (1982), pp. 144–237; and A.N. ZHUKOVA, *IAzyk palanskikh koriakov* (1980), a grammar of a southern Koryak dialect with texts, Russian translations, and dictionary appendices. (R.Au./D.Ab.)

Caucasian languages. *General works:* G. DEETERS, G.R. SOLTA, and VAHAN INGLISIAN, *Armenisch und kaukasische Sprachen* (1963), a survey, includes a presentation of the structure of the Caucasian languages according to the most characteristic features of phonology, morphology, and syntax, with an extensive bibliography. G.A. KLIMOV, *Kavkazskie iazyki* (1965), also available in a German translation, *Die kaukasischen Sprachen* (1969), offers a brief exposition of the history and structures of the Caucasian languages, with a general characterization of each group and an extensive bibliography, and his *Vvedenie v kavkazskoe iazykoznanie,* ed. by B.A. SEREBRENNIKOV (1986), also available in a German translation, *Einführung in die kaukasische Sprachwissenschaft,* ed. and trans. by JOST GIPPERT (1994), is also of interest. A.H. KUIPERS, "Caucasian," in *Current Trends in Linguistics,* vol. 1 (1963), pp. 315–344, provides a useful brief survey of Caucasian linguistics. ADOLF DIRR, *Einführung in das Studium der kaukasischen Sprachen* (1928, reprinted 1978), contains a survey of the structure of individual Caucasian languages and their interrelationships as well as a linguistic atlas. Much of the more recent scholarship is published in Georgian or Russian. Recent works in English include JOHN A.C. GREPPIN (ed.), *The Indigenous Languages of the Caucasus* (1991–); and HOWARD I. ARONSON (ed.), *Non-Slavic Languages of the U.S.S.R.: Papers from the Fourth Conference* (1994). (Th.V.G./T.E.G.)

South Caucasian languages: Two useful practical guides to Georgian are KITA TSCHENKÉLI, *Einführung in die georgische Sprache,* 2 vol. (1958, reissued 1980); and HOWARD I. ARONSON, *Georgian: A Reading Grammar,* corrected ed. (1990). Much useful information about Georgian and the history of Georgian (Kartvelian) studies is contained in SHOTA DZIDZIGURI (ŠOTʿA ŽIŽIGURI), *The Georgian Language* (1969). HANS VOGT, *Grammaire de la langue géorgienne* (1971), is a useful grammar. Important special works on the South Caucasian languages include G. DEETERS, *Das kharthwelische Verbum* (1930); A.S. ČIKOBAVA, *Drevnejšaja struktura imennyx osnov v kartvel'skix jazykax* (1942); KARL HORST SCHMIDT, *Studien zur Rekonstruktion des Lautstandes der südkaukasischen Grundsprache* (1962); T.V. GAMKRELIDZE and G.I. MACHAVARIANI, *Sonantṫa sistema da ablauti kʿartʿvelur enebši* (1965), in Georgian with a Russian summary, also available in a German translation, *Sonantensystem und Ablaut in den Kartwelsprachen* (1982); and ALICE C. HARRIS, *Diachronic Syntax: The Kartvelian Case* (1985). A.G. ŠANIDZE, *Osnovy gruzinskoj grammatiki,* vol. 1 (1953), is a comprehensive exposition (in Georgian) of the structure of modern Georgian. An account of the Georgian sound system is given in G.S. AXVLEDIANI, *Osnovy obščej fonetiki* (1949), in Georgian. A detailed descriptive analysis of the Svan verb system according to dialects is found in V.T. TOPURIA, *Svanuri ena,* vol. 1, *Zmna* (1931), in Georgian with a Russian summary. Additional works include DEE ANN HOLISKY, *Aspect and Georgian Medial Verbs* (1981); and T.V. GAMKRELIDZE, *Alphabetic Writing and the Old Georgian Script* (1994; originally published in Georgian, 1989). (Th.V.G./Ed.)

North Caucasian languages: Studies of various aspects of the North Caucasian languages include A. TSCHIKOBAVA, "Die ibero-kaukasischen Gebirgssprachen und der heutige Stand ihrer Erforschung in Georgien," *Acta Orientalia Academiae Scientiarum Hungaricae,* 9:109–161 (1959); N. TRUBETZKOY, "Nordkaukasische Wortgleichungen," *Wiener Zeitschrift für die Kunde des Morgenlandes,* vol. 37 (1930), pp. 79–92; GEORGES DUMÉZIL, *Études comparatives sur les langues caucasiennes du nord-ouest (morphologie)* (1932); A.H. KUIPERS, *Phoneme and Morpheme in Kabardian (East Adyghe)* (1960); W.S. ALLEN, "Structure and System in the Abaza Verbal Complex," *Transactions of the Philological Society* (1956), pp. 127–176; two essays by ALF SOMMERFELT in *Norsk Tidsskrift for Sprogvidenskap,* "Études comparatives sur le caucasique du Nord-ouest," 7:178–210 (1934), and "Études comparatives . . . Nord-est," 9:115–143 (1938); HENRICUS JOANNES SMEETS, *Studies in West Circassian Phonology and Morphology* (1984); JOHN COLARUSSO, *The Northwest Caucasian Languages: A Phonological Survey* (1988), and *A Grammar of the Kabardian Languages* (1992); and MARTIN HASPELMATH, *A Grammar of Lezgian* (1993). (T.E.G.)

Afro-Asiatic (Hamito-Semitic) languages. *General works:* There is still no general survey of the field (including bibliography) to replace I.M. DIAKONOFF, *Semito-Hamitic Languages* (1965; originally published in Russian, 1965); however, the developments in the field after 1965 have been considerable, as evidenced by, for example, JAMES BYNON and THEODORA BYNON (eds.), *Hamito-Semitica* (1975), papers from a colloquium. Among other general surveys are G.R. CASTELLINO, *The Akkadian Personal Pronouns and Verbal System in the Light of Semitic and Hamitic* (1962), and T.W. THACKER, *The Relationship of the Semitic and Egyptian Verbal Systems* (1954).

The theoretical problems of the Hamito-Semitic family have, until recently, been studied mostly on the basis of the Semitic branch alone. Especially important in this respect are I.J. GELB, *Sequential Reconstruction of Proto-Akkadian* (1969); JERZY KURYLOWICZ, *L'Apophonie en sémitique* (1961); and FRITHIOF RUNDGREN, *Intensiv und Aspektkorrelation* (1959). Perhaps the most crucial problem of the Proto-Hamito-Semitic linguistic typology is the reconstruction of the verbal system, for which, besides the above-mentioned works, see also MARCEL COHEN, *Le Système verbal sémitique et l'expression du temps* (1924); O. ROESSLER, "Akkadisches und libysches Verbum," *Orientalia,* vol. 20 (1951); A. KLINGENHEBEN, "Die Präfix- und die Suffixkonjugation des Hamitosemitischen," *Mitteilungen des Instituts für Orientforschung,* vol. 2 (1957); and the critical, although certainly not final, review of some later ideas in PELIO FRONZAROLI, "Ricostruzione interna del verbo semitico in alcuni studi recenti," *Accademia Toscana "La Colombaria,"* pp. 71–85 (1972). Another theoretical problem is broached in I.M. DIAKONOFF, "Problems of Root Structure in Proto-Semitic," *Archiv Orientální,* 38:453–480 (1970). The best general review of the Semitic branch of Hamito-Semitic is GOTTHELF BERGSTRAESSER, *Einführung in die semitischen Sprachen* (1928, reprinted 1963). See also GEORGIO LEVI DELLA VIDA (ed.), *Semitic Linguistics: Present and Future* (1961); SABATINO MOSCATI, *An Introduction to the Comparative Grammar of the Semitic Languages* (1964); and I.M. DIAKONOFF, *Jazyki drevnej Perednej Azii* (1967), on languages of the ancient Middle East, including Semitic and a survey of Common Hamito-Semitic. Recent developments are evaluated in THOMAS A. SEBEOK (ed.), *Current Trends in Linguistics,* vol. 6, *Linguistics in South West Asia and North Africa* (1970), with a comprehensive bibliography. JOSHUA BLAU, *The Renaissance of Modern Hebrew and Modern Standard Arabic* (1981), is a comparative study.

Akkadian: WOLFRAM VON SODEN, *Grundriss der akkadischen Grammatik,* 2nd ed. (1969), and *Akkadisches Handwörterbuch* (1959–); ERICA REINER, *A Linguistic Analysis of Akkadian* (1966); *The Assyrian Dictionary,* published by the Oriental Institute, the University of Chicago (1956–).

Northern Central Semitic: C.H. GORDON, *Ugaritic Textbook,* 2nd ed. (1965); JOSEPH AISTLEITNER, *Wörterbuch der ugaritischen Sprache* (1963); I.J. GELB, "La lingua degli Amoriti," *Rendiconti d. Accademia Nazionale dei Lincei,* 13:143–164 (1958); GIOVANNI GARBINI, *Il Semitico di Nord-Ovest* (1960); Z.S. HARRIS, *Development of the Canaanite Dialects* (1939); JOHANNES FRIEDRICH, *Phönizisch-punische Grammatik* (1951); GEORG BEER and RUDOLF MEYER, *Hebräische Grammatik,* 2 vol. (1952–55); WILHELM GESENIUS and GOTTHELF BERGSTRAESSER, *Hebräische Grammatik,* 29th ed., 2 vol. (1918–29); HANS BAUER and PONTUS LEANDER, *Historische Grammatik der Hebräischen Sprache des Alten Testamentes* (1922); H.B. ROSEN, *A Textbook of Israeli Hebrew,* 2nd ed. (1966); LUDWIG KOEHLER and WALTER BAUMGARTNER, *Lexicon in Veteris Testamenti Libros* (1958);

E. BEN YEHUDA, *Thesaurus totius Hebraitatis* (1908–58); FRANZ ROSENTHAL, *Die aramaistische Forschung seit Th. Nöldeke's Veröffentlichungen* (1939), and *A Grammar of Biblical Aramaic* (1961); PONTUS LEANDER, *Lautund Formenlehre des Ägyptisch-Aramäischen* (1928, reprinted 1966); JEAN CANTINEAU, *Le Nabatéen*, 2 vol. (1930–32); HARRIS BIRKELAND, *The Language of Jesus* (1954); E.Y. KUTSCHER, *Studies in Galilean Aramaic* (1952); THEODOR NOELDEKE, *Kurzgefasste syrische Grammatik* (1880, reprinted 1966); CARL BROCKELMANN, *Syrische Grammatik*, 8th ed. (1960); RUDOLF MACUCH, *Handbook of Classical and Modern Mandaic* (1965); KONSTANTIN CERETELI, "Abriss der vergleichenden Phonetik der modernen assyrischen Dialekte," in FRANZ ALTHEIM, *Geschichte der Hunnen*, vol. 3 (1961), pp. 218–266; IRENE GARBELL, *The Jewish Neo-Aramaic Dialect of Persian Azerbaijan* (1965); CHARLES F. JEAN and JACOB HOFTIJZER, *Dictionnaire des inscriptions sémitiques de l'Ouest* (1960–65); MARCUS JASTROW, *A Dictionary of the Targumim, The Talmud Babli and Jerushalmi, and the Midrashic Literature*, 2 vol. (1950); ROBERT PAYNE SMITH, *Thesaurus Syriacus*, 2 vol. (1868–1901); E.S. DROWER and RUDOLF MACUCH, *A Mandaic Dictionary* (1963).

Southern Central Semitic, or Arabic: CHAIM RABIN, *Ancient West Arabian* (1951); HENRI FLEISCH, *L'Arabe classique* (1956); JEAN CANTINEAU, *Cours de phonétique arabe* (1960) *and La Dialectologie arabe* (1955); A. SUTCLIFFE, *Grammar of the Maltese Language* (1936); E.W. LANE, *An Arabic-English Lexicon*, 8 vol. (1863–93).

Southern Peripheral Semitic: A.F.L. BEESTON, *A Descriptive Grammar of Epigraphic South Arabian* (1962); MARIA HOEFNER, *Altsüdarabische Grammatik* (1943); EWALD WAGNER, *Syntax der Mehri-Sprache* (1953); WOLF LESLAU, *Lexique soqoṭri* (1938); M. BITTNER, "Charakteristik der Sprache der Insel Soqotra," *Anzeiger der Wiener Akademie der Wissenschaften, Ph.-hist. Kl.*, vol. 55 (1918)—the same author has published a number of studies on the important languages Mahrī and Shaḥrī ("Shkhauri") in the *Sitzungsberichte* of the *Wiener Akademie* between 1909 and 1917; CHRISTIAN DILLMANN, *Ethiopic Grammar*, 2nd ed. (1907), and *Lexicon linguae Aethiopicae* (1865); WOLF LESLAU, *Étude descriptive et comparative du Gafat* (1956) and *Etymological Dictionary of Harari* (1963); EDWARD ULLENDORFF, *The Semitic Languages of Ethiopia: A Comparative Phonology* (1955).

Egyptian: ELMAR EDEL, *Altägyptische Grammatik* (1955–64); A.H. GARDINER, *Egyptian Grammar*, 3rd ed. (1957); WALTER TILL, *Koptische Grammatik (Saïdischer Dialekt)* (1955); ADOLF ERMAN and HERMANN GRAPOW (eds.), *Wörterbuch der ägyptischen Sprache*, 6 vol. (1926–31, reprinted 1955); WOLJA ERICHSEN, *Demotisches Glossar* (1954).

Berber: Two works by ANDRE BASSET, *Handbook of African Languages*, vol. I, *La Langue berbère* (1952), and *Articles de dialectologie berbère* (1959), are useful introductions. The article by JOSEPH R. APPLEGATE, "The Berber Languages," in THOMAS A. SEBEOK (ed.), *Current Trends in Linguistics*, vol. 6 (1970), pp. 586–608, is followed by an extensive classified bibliography, pp. 608–661.

Cushitic: M.L. BENDER, "The Languages of Ethiopia: A New Lexicostatistic Classification and Some Problems of Diffusion," *Anthropological Linguistics*, vol. 13 (1971); A.B. DOLGOPOL'SKIJ, *Sravitel'no-istoricheskaja fonetika kušitskikh jazykov* (1973); on the individual branches and languages of Cushitic, see C.R. BELL, *The Somali Language* (1953), a manual of the Isāq dialect; M.M. MORENO, *Il somalo della Somalia* (1955), devoted to the Benadir, Darod, and Digil dialects; ENRICO CERULLI, *Studi etiopici*, 4 vol. (1936–51), contains grammars and vocabularies of Sidamo, Janjero, some Ometo dialects, and Kafa (Kaficho). For a good survey of the individual branches and languages of Cushitic, see F.R. PALMER, "Cushitic," in THOMAS A. SEBEOK (ed.), *Current Trends in Linguistics*, vol. 6 (1970), pp. 571–585.

Chadic: On the general problems of the Chadic branch see PAUL NEWMAN and ROXANA MA, "Comparative Chadic: Phonology and Lexicon," *Journal of African Languages*, vol. 5 (1966); D. WESTERMANN and M. BRYAN, "Languages of West Africa," *Handbook of African Languages*, vol. 2 (1952); JOSEPH H. GREENBERG, "Studies in African Linguistic Classification, IV, Hamito-Semitic," *SWest. J. Anthrop.*, vol. 6 (1950). Among the studies of individual Chadic languages those devoted to Hausa are the most numerous of all; the following may be especially useful: C.T. HODGE and IBRAHIM UMARU, *Hausa: Basic Course* (1963); CHARLES H. KRAFT, *A Study of Hausa Syntax*, 3 vol. (1963); and R.C. ABRAHAM, *The Language of the Hausa People* (1959). Among the grammatical studies of the other Chadic languages are HERRMANN JUNGRAITHMAYR, *Die Ron-Sprachen* (1970); CARL HOFFMAN, *A Grammar of the Margi Language* (1963); and H.D. FOULKES, *Angass Manual* (1915). Of the many vocabularies of Hausa the most important is G.P.A. BARGERY (comp.), *A Hausa-English Dictionary and English-Hausa Vocabulary* (1934). (I.M.D.)

Korean language. SAMUEL E. MARTIN, *A Reference Grammar of Korean* (1992), provides a comprehensive description. HO-MIN SOHN, *Korean* (1994), is a detailed grammar. Textbooks include FRED LUKOFF, *Spoken Korean*, 2 vol. (1945–47); B. NAM PARK, *Korean Basic Course*, 2 vol. (1968–69); SAMUEL E. MARTIN and YOUNG-SOOK C. LEE, *Beginning Korean* (1969, reissued 1986); and two series of textbooks published by Korea University (Seoul): *Han'gugŏ*, 4 vol. in various editions (1989–91); and *Han'gugŏ Hoehwa*, 4 vol. in various editions (1991–93). Informative articles may be found in *The Korean Language* (1983), edited by the Korean National Commission for UNESCO. The writing system is thoroughly discussed in YOUNG-KEY KIM-RENAUD (ed.), *The Korean Alphabet: Its History and Structure* (1997). Transcription is discussed in G.M. MCCUNE and E.O. REISCHAUER, "The Romanization of the Korean Language," *Transactions of the Korea Branch of the Royal Asiatic Society*, 29:1–55 (1939), with an abbreviated account in the same journal, "Tables of the McCune-Reischauer System for the Romanization of Korean," 38:121–129 (1961); and in ROBERT AUSTERLITZ *et al.*, "Report of the Workshop Conference on Korean Romanization," *Korean Studies*, 4:111–125 (1980). S. ROBERT RAMSEY, *Accent and Morphology in Korean Dialects* (1978), is a valuable explanation of accentual patterns. The history of the language is chronicled in KI-MUN YI, *Kugŏsa kaesŏl*, 2nd ed. (1972), also available in a German edition, *Geschichte der koreanischen Sprache*, trans. and ed. by BRUNO LEWIN (1977). The prehistory is reviewed in HO-MIN SOHN, "The State of the Art in the Historical-Comparative Studies of Japanese and Korean," *Korean Studies*, 4:29–50 (1980); and further explored in three works by SAMUEL E. MARTIN: "Morphological Clues to the Relationships of Japanese and Korean," in PHILIP BALDI (ed.), *Linguistic Change and Reconstruction Methodology* (1990), pp. 483–509, "On the Prehistory of Korean Grammar: Verb Forms," *Korean Studies*, 19:139–150 (1995), and *Consonant Lenition in Korean and the Macro-Altaic Question* (1996). The authoritative dictionary is *Uri mal kŭn sajŏn*, 4 vol. (1991–92), compiled by Han'gŭl Hakhoe (Korean Language Research Society). Also very useful is the dictionary by YI SUNG-NYŎNG (SUNG-NYONG LEE), *Kugŏ taesajŏn*, 3 vol. (1960); and the North Korean dictionary *Chosŏn mal sajŏn*, 6 vol. (1960–62). The recommended Korean-English dictionary is SAMUEL E. MARTIN, YANG HA LEE (YANG-HA YI), and SUNG-UN CHANG (SŎNG-ON CHANG), *A Korean-English Dictionary* (1967). Y.H. LEE and J.W. KWUN (eds.), *Minjungsugwan's Pocket English-Korean Dictionary*, new rev. ed., 2 vol. (1960), is useful for translating English words. (S.E.M.)

Japanese language. Two widely used college textbooks are OSAMU MIZUTANI and NOBUKO MIZUTANI, *An Introduction to Modern Japanese* (1977); and ELEANOR HARZ JORDEN and MARI NODA, *Japanese, the Spoken Language*, 3 vol. (1987–90). ROY ANDREW MILLER, *The Japanese Language* (1967, reprinted 1980), provides a general introduction with emphasis on historical development, and his *Japanese and the Other Altaic Languages* (1971), documents the hypothesis relating Japanese to the Altaic family. G.B. SANSOM, *An Historical Grammar of Japanese* (1928, reissued 1995), is an excellent introduction to Old Japanese, detailing the development of the writing systems. A historical survey of the literary language can be found in YAEKO SATO HABEIN, *The History of the Japanese Written Language* (1984). SAMUEL E. MARTIN, *A Reference Grammar of Japanese* (1975, reprinted with corrections, 1988), is a comprehensive description of Modern Japanese, with numerous examples taken from published materials, while his *The Japanese Language Through Time* (1987), traces the historical developments of major grammatical structures. SUSUMU KUNO, *The Structure of the Japanese Language* (1973), is a particularly useful introduction, emphasizing particles and topic construction. JAMES D. MCCAWLEY, *The Phonological Component of a Grammar of Japanese* (1968), technically treats Japanese phonology and includes a useful survey of Japanese-dialect accentual systems. A recent comprehensive survey encompassing all the major fields of Japanese linguistics is MASAYOSHI SHIBATANI, *The Languages of Japan* (1990). (M.Si.)

Austronesian languages. A classic descriptive study is HERMANUS NEUBRONNER VAN DER TUUK, *A Grammar of Toba Batak* (1971; originally published in Dutch, 2 vol., 1864–67). RENWARD BRANDSTETTER, *An Introduction to Indonesian Linguistics*, trans. by C.O. BLAGDEN (1916), provides a still readable and informative overview of the Austronesian languages of island Southeast Asia and introduces the concept of the submorphemic element he called a *Wurzel* or "root." The foundation for all subsequent comparative work is OTTO DEMPWOLFF, *Vergleichende Lautlehre des austronesischen Wortschatzes*, 3 vol. (1934–38, reprinted 1969); neither of the English translations of this work is to be trusted. The major modifications of Dempwolff's results appear in ISIDORE DYEN, *The Proto-Malayo-Polynesian Laryngeals* (1953); OTTO C. DAHL, *Proto-Austronesian*, 2nd rev. ed. (1977); ROBERT A. BLUST, *The Proto-*

Oceanic Palatals (1978); and M.D. ROSS, *Proto Oceanic and the Austronesian Languages of Western Melanesia* (1988), and "The Sound of Proto-Austronesian: An Outsider's View of the Formosan Evidence," *Oceanic Linguistics,* 31(1):23–64 (Summer 1992). The most important discussions of the linguistic position of Malagasy are OTTO C. DAHL, *Malgache et Maanjan: un comparaison linguistique* (1951); and K.A. ADELLAR, "Malay Influence on Malagasy: Linguistic and Culture-Historical Implications," *Oceanic Linguistics,* 28(1):1–46 (Summer 1989). A.K. PAWLEY, "Polynesian Languages: A Subgrouping Based on Shared Innovations in Morphology," *The Journal of the Polynesian Society,* 75:39–64 (1966), laid the foundations for the internal subgrouping of Polynesian languages. Brandstetter's theory of the root is examined in greater detail in ROBERT A. BLUST, *Austronesian Root Theory: An Essay on the Limits of Morphology* (1988). Linguistic reconstruction in a number of semantic fields is explored in A.K. PAWLEY and M.D. ROSS (eds.), *Austronesian Terminologies: Continuity and Change* (1994); and contact-induced change in Austronesian languages is treated at length in TOM DUTTON and DARRELL TRYON (eds.), *Language Contact and Change in the Austronesian World* (1994). Two major Festschriften with numerous papers on Austronesian languages are DONALD C. LAYCOCK and WERNER WINTER (eds.), *A World of Language* (1987); and ROBERT A. BLUST (ed.), *Currents in Pacific Linguistics* (1991). DARRELL T. TRYON (ed.), *Comparative Austronesian Dictionary: An Introduction to Austronesian Studies,* 4 vol. (1995), although misnamed, contains the most complete list of Austronesian languages together with many useful sketches of phonology, morphology, and grammar. The most important studies on the possible external relationships of Austronesian languages are PAUL K. BENEDICT, *Austro-Thai Language and Culture, with a Glossary of Roots* (1975); and LAWRENCE A. REID, "Morphological Evidence for Austric," *Oceanic Linguistics,* 33(2):323–344 (December 1994). (R.A.Bl.)

Papuan languages. Most earlier general studies have become obsolete. Comprehensive recent bibliographical and factual information may be found in two essays in THOMAS A. SEBEOK (ed.), *Current Trends in Linguistics,* vol. 8, part 1 (1971): D.C. LAYCOCK and C.L. VOORHOEVE, "History of Research in Papuan Languages," pp. 509–540; and STEPHEN A. WURM, "The Papuan Linguistic Situation," pp. 541–657—Wurm's essay is updated in his book *Papuan Languages of Oceania* (1972). Individual language descriptions are given in H. MCKAUGHAN (ed.), *Languages of the Eastern Family* (1971). Extensive descriptive and comparative materials on Papuan languages appear in the serial publications *Pacific Linguistics.*

Australian Aboriginal languages. A concise discussion of all aspects of the study of Australian Aboriginal languages, with extensive bibliography, is provided in S.A. WURM, *Languages of Australia and Tasmania* (1972); extensive information may be found in three essays in THOMAS A. SEBEOK (ed.), *Current Trends in Linguistics,* vol. 8, part 1 (1971): A. CAPELL, "History of Research in Australian and Tasmanian Languages," pp. 661–720; STEPHEN A. WURM, "Classifications of Australian Languages, Including Tasmanian," pp. 721–778; and GEOFFREY N. O'GRADY, "Lexicographic Research in Aboriginal Australia," pp. 779–803; and in GEOFFREY N. O'GRADY and C.F. and F.M. VOEGELIN, *Classification and Index of the World's Languages* (1977). Numerous articles and monographs on Australian languages appear in the serials *Australian Aboriginal Studies,* Canberra; *Pacific Linguistics,* Canberra; *Oceania,* Sydney; *Oceania Linguistic Monographs,* Sydney. See also the overview in BARRY J. BLAKE, *Australian Aboriginal Languages* (1981). (S.A.W.)

African languages. *General works:* ISTVAN FODOR, *The Problems in the Classification of the African Languages: Methodological and Theoretical Conclusions Concerning the Classification System of Joseph H. Greenberg* (1959); JOSEPH H. GREENBERG, *Studies in African Linguistic Classification* (1955), Greenberg's first classification of African languages; "Africa As a Linguistic Area," in W.R. BASCOM and M.J. HERSKOVITS (eds.), *Continuity and Change in African Cultures,* pp. 15–27 (1959); *The Languages of Africa* (1963), a revision of Greenberg (1955), representing the most up-to-date version of the author's classification; HARRY H. JOHNSTON, *A Comparative Study of the Bantu and Semi-Bantu Languages,* 2 vol. (1919–22), a pioneering study showing many significant similarities between Bantu and many West African languages, both in structure and vocabulary; RICHARD LEPSIUS, *Nubische Grammatik, mit einer Einleitung über die Völker und Sprachen Afrika's* (1880), a grammar of Nubian with an extensive introduction containing one of the earliest overall classifications of African languages; CARL MEINHOF, *Die Sprachen der Hamiten* (1912), discusses a selection of languages considered by the author to be Hamitic, including Masai, and the linguistic characteristics that link them; FRIEDRICH MUELLER, *Grundriss der Sprachwissenschaft,* 4 vol. (1876–88), contains one of the earliest attempts at an overall classification of African languages; GEORGE PETER MURDOCK, *Africa: Its Peoples and Their Culture History* (1959); THOMAS A. SEBEOK (ed.), *Current Trends in Linguistics,* vol. 7, *Linguistics in Sub-Saharan Africa* (1971); ARCHIBALD N. TUCKER and MARGARET A. BRYAN, *The Non-Bantu Languages of North-Eastern Africa* (1956), the most detailed survey of all the Chari-Nile languages and language groups, plus many others; *Linguistic Analyses: The Non-Bantu Languages of North-Eastern Africa* (1966), linguistic sketches of the languages included in Tucker and Bryan (1956); DIEDRICH WESTERMANN, *Die Sudansprachen, eine sprachvergleichende Studie* (1911), an attempt to demonstrate the interrelationship of five West African languages (Twi, Ga, Ewe, Yoruba, and Efik) and three from East Africa (Kunama, Nubian, and Dinka), thereby seeking to show the unity of all the Sudanic languages; *Die westlichen Sudansprachen und ihre Beziehungen zum Bantu* (1927), weakens the case for the interrelationship of all the Sudanic languages by showing a closer interrelationship between the Western Sudanic and Bantu languages; "Charakter und Einteilung der Sudansprachen," *Africa,* 8:129–148 (1935), revises the author's earlier views on the Sudanic languages; and with MARGARET A. BRYAN, *The Languages of West Africa* (1952), the most detailed survey of the languages of West Africa. Also useful is MELVIN K. HENDRIX, *An International Bibliography of African Lexicons* (1982). (M.F.Go.)

Niger-Congo languages: THOMAS A. SEBEOK (ed.), *Current Trends in Linguistics,* vol. 7, *Linguistics in Sub-Saharan Africa* (1971), is the most authoritative source on the Niger-Congo languages. A summary of selected facts about the languages may be found in the series called *Handbook of African Languages,* sponsored by the International African Institute, especially the two volumes on Bantu by C.M. DOKE (1945 and 1954), the two on the same subject by MALCOLM GUTHRIE (1948 and 1953), and the one by DIEDRICH WESTERMANN and MARGARET A. BRYAN on the *Languages of West Africa* (*op. cit.*). The most important comparative study of the Bantu languages is presented in MALCOLM GUTHRIE, *Comparative Bantu: An Introduction to the Comparative Linguistics and Prehistory of the Bantu Languages,* 4 vol. (1967–70). The following periodicals are devoted entirely to articles on African languages, of which the overwhelming majority are concerned with Niger-Congo: *Bantu Studies; Journal of African Languages; Journal of West African Languages; Afrika und Übersee; Studies in African Linguistics; African Language Studies;* and *African Language Review.* (D.W.C.)

Chari-Nile and Nilo-Saharan languages: MERVYN W.H. BEECH, *The Suk: Their Language and Folklore,* with introduction by SIR CHARLES ELIOT (1911); MARGARET A. BRYAN, "The T/K Languages: A New Substratum," *Africa,* 29:1–21 (1959); "The *N/*K Languages of Africa," *J. Afr. Lang.,* 7:169–217 (1968); MORRIS GOODMAN, "Some Questions on the Classification of African Languages," *Int. J. Am. Linguistics,* 36:117–122 (1970); OSWIN KÖHLER, "Geschichte der Erforschung der Nilotischen Sprachen," *Afrika und Übersee,* vol. 28 (1955), a history of the investigation of the Nilotic languages, including "Nilo-Hamitic," and suggesting a new subgrouping; G.W. MURRAY, "The Nilotic Languages: A Comparative Survey," *The Journal of the Royal Anthropological Institute of Great Britain and Ireland,* 50:327–368 (1920), an attempt to prove the common ancestry of Nubian, Bari, Masai, and Shilluk peoples; P.L. SHINNIE, *Meroe: A Civilization of the Sudan* (1967), includes a discussion of the Meroitic language; BRUCE G. TRIGGER, "Meroitic and Eastern Sudanic: A Linguistic Relationship?," *Kush,* 12:188–194 (1964); ARCHIBALD N. TUCKER, *The Eastern Sudanic Languages* (1940), compares, among others, the languages renamed Central Sudanic by Greenberg and concentrates on the Moru-Madi group and Lendu; DIEDRICH WESTERMANN, *The Shilluk People: Their Language and Folklore* (1912), discusses the classification of the Nilotic languages. (M.F.Go.)

Khoisan languages: (*Classification*): E.O.J. WESTPHAL, "A Re-classification of Southern African Non-Bantu Languages," *J. Afr. Lang.,* 1:1–8 (1962), "The Click Languages of Southern and Eastern Africa," in THOMAS A. SEBEOK (ed.), *Current Trends in Linguistics,* vol. 7 (1971), pp. 367–420, and "The Linguistic Prehistory of Southern Africa: Bush, Kwadi, Hottentot, and Bantu Linguistic Relationships," *Africa,* 33:237–265 (1963), with maps; L.W. LANHAM and D.P. HALLOWES, "Linguistic Relationships and Contacts Expressed in the Vocabulary of Eastern Bushman," *Afr. Stud.,* 15:45–48 (1956); for a comprehensive list of classified languages, see D.F. BLEEK, *A Bushman Dictionary* (1956). (*Distribution*): OSWIN KÖHLER, "Die Khoe-sprachigen Buschmänner der Kalahari: Ihre Verbreitung und Gliederung," *Kölner Geogr. Arbeiten. Festschrift Kurt Kayser,* pp. 373–411 (1971), with map. (*Phonology*): C.M. DOKE, "An Outline of the Phonetics of the Language of the Chũ: Bushmen of North-West Kalahari," *Bantu Stud.,* 2:129–166 (1925); D.M. BEACH, *The Phonetics of the Hottentot Language* (1938). (*Grammar*): OSWIN KÖHLER, "Noun Classes and Grammatical Agreement in !Xũ," *Actes du VIII^e Congrès Intern. de Linguistique Africaine,* pp. 489–522 (1971); J.W. SNYMAN, *An Introduction to the !Xu (!Kung) Language* (1970); D.F. BLEEK, "Bushman Grammar: A Grammatical Sketch of the Language of the |xam-ka-!k'e'," *Zeitschrift für Eingeborenen-Sprachen,* vol. 19–20 (1928–30); L.F. MAINGARD, "The ≠Khomani Dialect of Bushman," in J.D.R. JONES and C.M. DOKE (eds.), *Bushmen of the Southern Kalahari* (1937); F. RUST, *Praktische Namagrammatik* (1965), based on material of H. Vedder and J. Olpp. (*Comparative studies*): OSWIN KÖHLER, "Studien zum Genussystem und Verbalbau der zentralen Khoisan-Sprachen," *Anthropos,* 57:529–546 (1962), deals with gender and verb system of Central Khoisan; D.F. BLEEK, "A Short Survey of Bushman Languages," *Zeitschrift für Eingeborenen-Sprachen,* vol. 30 (1939–40). (*Vocabulary*): D.F. BLEEK, *A Bushman Dictionary* (1956), a compilation of all language material available, except Khoikhoin; J.G. KRÖNLEIN, *Wortschatz der Khoi-Khoin-Namaqua-Hottentoten* (1889), on Nama. (*General information*): I. SCHAPERA, *The Khoisan Peoples of South Africa* (1930), deals also with distribution and classification. (*Eyasian languages*): O. DEMPWOLFF, *Die Sandawe* (1916), with a chapter on language and vocabulary; D.F. BLEEK, "Traces of Former Bushman Occupation in Tanganyika Territory," *S. Afr. J. Sci.,* 28:423–429 (1931). (D.R.A.K.)

Eskimo-Aleut languages. Bibliography, demographic data, and literacy programs are addressed in DIRMID R.F. COLLIS (ed.), *Arctic Languages: An Awakening* (1990). Linguistic information may be found in ANTHONY C. WOODBURY, "Eskimo and Aleut Languages," in WILLIAM C. STURTEVANT (ed.), *Handbook of North American Indians,* vol. 5, *Arctic,* ed. by DAVID DAMAS (1984), pp. 49–63; MICHAEL FORTESCUE, STEVEN JACOBSON, and LAWRENCE KAPLAN, *Comparative Eskimo Dictionary with Aleut Cognates* (1994); MICHAEL FORTESCUE, *West Greenlandic* (1984); STEVEN JACOBSON and ANNA W. JACOBSON, *A Practical Grammar of the Central Alaskan Yup'ik Eskimo Language* (1995); JOSEPH DE REUSE, *Siberian Yupik Eskimo: The Language and Its Contacts with Chukchi* (1994); STEVEN A. JACOBSON (compiler), *Yup'ik Eskimo Dictionary* (1984); and KNUT BERGSLAND (compiler), *Aleut Dictionary* (1994). (K.B.)

North American Indian languages. J.W. POWELL, "Indian Linguistic Families of America North of Mexico," *U.S. Bureau of American Ethnology, 7th Annual Report* (1891), pp. 1–142, the first comprehensive classification; FRANZ BOAS, *Handbook of American Indian Languages,* 3 pt. (1911, 1922, and 1933–38), a classic introduction, with sketches of sample languages; HARRY HOIJER *et al., Linguistic Structures of Native America* (1946), a summary of work on language classification and sketches of languages; C.F. and F.M. VOEGELIN, *Map of North American Indian Languages,* rev. ed. (1966), presents the classification used in this article; THOMAS A. SEBEOK (ed.), *Current Trends in Linguistics,* vol. 10, *North America* (1973), surveys of different aspects of the field, with extensive bibliographies (see especially the valuable article of JOEL SHERZER, "Areal Linguistics in North America"); M.R. HAAS, *The Prehistory of Languages* (1969), a discussion of the principles in the historical study of American Indian languages; WALLACE CHAFE, "Estimates Regarding the Present Speakers of North American Indian Languages," *International Journal of American Linguistics,* 28:162–171 (1962), and 31:345–346 (1965), gives data used in the present article; EDWARD SAPIR, *Selected Writings in Language, Culture, and Personality* (1949), articles on the relationship of language and culture in aboriginal North America; B.L. WHORF, *Language, Thought, and Reality: Selected Writings* (1956), classic articles on American Indian language and worldview. (W.O.B.)

Meso-American Indian languages. Few books have appeared treating the Meso-American Indian languages as a group, although many dictionaries and grammars for individual languages have been prepared. An overview can be gotten from three articles in the *Handbook of Middle American Indians,* vol. 5, *Linguistics* (1967): MARIA TERESA FERNANDEZ DE MIRANDA, "Inventory of Classificatory Materials," an annotated bibliography; MORRIS SWADESH, "Lexicostatistic Classification," in which the author applies glottochronology to the classification of all the Meso-American Indian languages; and ROBERT LONGACRE, "Systemic Comparison and Reconstruction," a review of what had been accomplished in the historical-comparative field to that date. For linguistic characteristics of Meso-American languages, see TERRENCE KAUFMAN, "Areal Linguistics and Middle America," in THOMAS A. SEBEOK (ed.), *Current Trends in Linguistics,* vol. 11 (1972), pp. 459–484. For Uto-Aztecan, see C.F. and F.M. VOEGELIN and KENNETH L. HALE, *Typological and Comparative Grammar of Uto-Aztecan* (1962); and WICK R. MILLER, *Uto-Aztecan Cognate Sets* (1967). For Mayan and "new languages," see TERRENCE KAUFMAN, "Teco—A New Mayan Language," *Int. J. Am. Linguistics,* 35:154–174 (1969). For "external contacts," see RONALD D. OLSON, "Mayan Affinities with Chipaya of Bolivia," *ibid.,* 30:313–324 (1964), and 31:29–38 (1965). (Te.K.)

South American Indian languages. CESTMIR LOUKOTKA, *Classification of South American Indian Languages* (1968), contains a full list of languages, an extensive and detailed bibliography, the location and classification of languages, and a map. J.A. MASON, "The Languages of South American Indians," in J.H. STEWARD (ed.), *Handbook of South American Indians,* vol. 6 (1950), pp. 157–317, has an extensive bibliography, general information on language groups, a discussion of classifications, and a map. Additional informative works include N.A. MCQUOWN, "The Indigenous Languages of Latin America," *Am. Anthrop.,* 57:501–570 (1955), a critical appraisal of classifications with useful lists of languages and families, the location and classification of languages, and a map; C.F. and F.M. VOEGELIN, "Languages of the World: Native America," *Anthrop. Linguistics,* vol. 6, no. 6, and vol. 7, no. 7 (1964–65), providing general information on American Indian languages, much information on groups and individual languages, and a discussion of classifications; THOMAS A. SEBEOK (ed.), *Current Trends in Linguistics,* vol. 4, *Ibero-American and Caribbean Linguistics,* pt. 2, "Linguistics of Non-Ibero-American Languages" (1968), critical surveys of work done during the previous 20 years; MORRIS SWADESH, "Afinidades de las lenguas amerindias," in *Akten des 34. Internationalen Amerikanisten Kongress,* pp. 729–738 (1964); and J.H. GREENBERG, "The General Classification of Central and South American Languages," in ANTHONY F.C. WALLACE (ed.), *Men and Cultures* (1959), two recent classifications of South American languages. (J.A.S.)

Sumerian language. ARNO POEBEL, *Grundzüge der sumerischen Grammatik* (1923), partly out of date, but still the only full grammar of Sumerian in all its stages; ADAM FALKENSTEIN, *Grammatik der Sprache Gudeas von Lagaš,* 2 vol. (1949–50), a very thorough grammar of the New Sumerian dialect, and *Das Sumerische* (1959), a very brief but comprehensive survey of the Sumerian language; CYRIL J. GADD, *Sumerian Reading Book* (1924), outdated but the only grammatical tool in English; SAMUEL N. KRAMER, *The Sumerians* (1963), provides a general introduction to Sumerian civilization. (I.J.G.)

Etruscan language. The following sources provide accurate and reliable information on Etruscan: M. PALLOTTINO, *Etruscologia,* 6th ed. (1968; Eng. trans., *The Etruscans,* 1955), the standard work; A.J. PFIFFIG, *Die etruskische Sprache* (1969), an excellent and complete statement of what is known about the Etruscan language; and MURRAY FOWLER and R.G. WOLFE, *Materials for the Study of the Etruscan Language,* 2 vol. (1965), which contains the inscriptions with indexes of several kinds for easy reference—there is no grammatical information included. (M.Fo.)

Basque language. RENE LAFON, "La lengua vasca," *Enciclopedia lingüística hispánica,* vol. 1 (1960), perhaps the best short introduction to Basque, both descriptive and historical; HUGO SCHUCHARDT, *Primitiae linguae Vasconum,* 2nd ed. (1968), detailed commentary of an Old Basque text; P. LAFITTE, *Grammaire basque,* 2nd ed. (1962), a standard normative grammar; J. COROMINES, *Estudis de toponímia catalana,* 2 vol. (1965–70), presents new data on the survival of Basque dialects in the Middle Ages; J.M. LACARRA, *Vasconia medieval* (1957), authoritative review by an historian of the linguistic situation in and around the Basque country; LUIS MICHELENA, *Fonética histórica vasca* (1961), essay on the reconstruction of the phonological system of Proto-Basque; LUIS MICHELENA (ed.), *Textos arcaicos vascos* (1964), an annotated collection of documents from antiquity to 1700; RENE LAFON, *Le Système du verbe basque au XVIe siècle,* 2 vol. (1943), the best account of form and function of the Basque verb; A. TOVAR, *La lengua vasca,* 2nd ed. (1954; abridged Eng. trans., *The Basque Language,* 1957), and *The Ancient Languages of Spain and Portugal* (1961), a discussion of the problem of the position of Basque among these now-extinct languages. (L.M.)

Pidgin. An extensive discussion and bibliography of the problems of pidgins and creoles is given in R.A. HALL, JR., *Pidgin and Creole Languages* (1966). A specialized grammar study is PIETER MUYSKEN (ed.), *Generative Studies on Creole Languages* (1981). See also HENRI TINELLI, *Creole Phonology* (1981). (R.A.H.)

The History of Latin America

Latin America is generally understood to consist of the entire continent of South America in addition to Mexico, Central America, and the islands of the Caribbean whose inhabitants speak a Romance language. The peoples of this large area shared the experience of conquest and colonization by the Spaniards and Portuguese from the late 15th through the 18th centuries as well as movements of independence from Spain and Portugal in the early 19th century. Even since independence, many of the various nations have experienced similar trends, and they have some awareness of a common heritage. However, there are also enormous differences between them. Not only do the people live in a large number of independent units, but the geography and climate of their countries vary immensely, and their social and cultural characteristics differ according to the different constitution of the inhabitants before the Iberian conquest and the different timing and nature of European occupation.

Since the Spanish and Portuguese element looms so large in the history of the region, it is sometimes proposed that Iberoamerica would be a better term than Latin America. Latin seems to suggest an equal importance of the French and Italian contributions, which is far from being the case. Nevertheless, usage has fastened on Latin America, and it is retained here.

This article treats the history of Latin America from the first occupation by Europeans to the late 20th century, with an initial consideration of the indigenous and Iberian background. For more detailed coverage of the area prior to European contact, see PRE-COLUMBIAN CIVILIZATIONS. For additional information about the European exploration and colonization of Latin America, see EUROPEAN OVERSEAS EXPLORATION AND EMPIRE, THE HISTORY OF. The nations of Central America are treated in the article CENTRAL AMERICA; there is also a separate Macropædia article MEXICO. For information about the individual countries of South America, see specific country articles by name—*e.g.,* BRAZIL, VENEZUELA, and ARGENTINA. The island nations of the Caribbean are discussed in the article WEST INDIES. The physical and human geography of the continents, with some historical overview, are provided in the articles NORTH AMERICA and SOUTH AMERICA. There is also a separate article LATIN-AMERICAN LITERATURE. For discussion of major cities of Latin America and their histories, see specific articles by name—*e.g.,* RIO DE JANEIRO, BUENOS AIRES, and MEXICO CITY.

For coverage of related topics in the *Macropædia* and *Micropædia,* see the *Propædia,* sections 951, 952, 964, 966, and 974, and the *Index.*

This article is divided into the following sections:

The background

Though the conditions of pre-Columbian America and 15th-century Iberia are beyond the scope of Latin-American history proper, they must be given consideration in that connection. Not only did the geography of precontact America persist, but both the new arrivals and the indigenous inhabitants long retained their respective general characteristics, and it was the fit between them that determined many aspects of Latin-American evolution.

THE INDIGENOUS WORLD AND THE WORD "INDIAN"

From the time of Columbus and the late 15th century forward, the Spaniards and Portuguese called the peoples of the Americas "Indians"—that is, inhabitants of India. Not only is the term erroneous by origin, but it did not correspond to anything in the minds of the indigenous people. They had no word meaning "inhabitant of the Western Hemisphere," and most of them seem not to have adopted any equivalent even after centuries of contact. Any such word refers to commonalities seen from the outside and not to any unity perceived by the inhabitants of the Americas themselves. The indigenous peoples were greatly varied, far more so than the Europeans; they were spread over a vast area and not aware of each other from one major region to the next.

Nonetheless, the indigenous peoples had several things in common. They were closely related to one another in biological terms, and their languages, though they cannot be shown to have a common origin, tend to share many general features. All shared an isolation from the great mass of humanity inhabiting Eurasia and Africa, who were in some way in contact with one another. The inhabitants of America all lacked immunities to diseases common in Europe and Africa. They had some impressive innovations to their credit, including the domesticated plants of Mesoamerica and the Andes, but all had been kept apart from things that had long since spread over much of the rest of the globe, including steel, firearms, horses, wheeled vehicles, long-distance shipping, and alphabetic writing. As a result the indigenous peoples, once in contact, were very vulnerable to the outsiders. Epidemics raged wherever intruders appeared; with their materials and techniques the Europeans were able to conquer whenever they felt it imperative to do so. There is, then, at times, a need for a common term, and if one realizes its limitations, "Indian" may do as well as another.

TYPES OF WESTERN HEMISPHERE SOCIETIES

The Europeans were sedentary, living in nations and districts with distinct borders, relying on a permanent intensive agriculture to sustain many people in a variety of pursuits who lived in both urban and rural communities. One large section of the indigenous American population, in fact the most numerous, based in Mesoamerica (central and southern Mexico and Guatemala) and the central Andes, was also sedentary. Indeed, these peoples and the Europeans tended to have more in common with each other than either had with the other peoples indigenous to the Americas. Another type of indigenous peoples may be called semisedentary. They lacked the permanent-site agriculture and the fixed borders of the sedentary peoples and were apparently far less numerous, but they had shifting agriculture and sizable, if frequently moving, settlements. They were found above all in relatively temperate forested areas. The third category that can be established is that of the nonsedentary peoples, who had little or no agriculture and moved annually in small bands over a large territory, hunting and gathering. They were located primarily in areas that under the then-existing technologies were not propitious for agriculture, especially plains and dense tropical forests.

Similarities with Europeans

Sedentary peoples. The sedentary peoples shared with the Europeans not only an agricultural base and dense, quite concentrated populations but also territorial states, hereditary rulers, state religions with priesthoods, specialized craft groups, social classes including a nobility distinct from commoners, and regularized taxes or tributes. Among some sedentary groups, large political structures—confederations or empires—had come into existence, collecting tribute and engaging in trade over long distances. The most famous of these are the Inca empire in the Andean region and what is often called the Aztec empire in Mexico (although the word Aztec was little known at the time). These empires were not nations but had at their centre one small ethnic state (or a few) that exercised dominance over a large number of similar states. The subject states retained their ethnic identity, their own rulerships, and their general way of life despite owing tribute to the imperial power. It was these subject entities that were to survive the conquest and serve as the base of the European presence. They had different names in different places, and indeed their structures varied, but they were everywhere enough like European small principalities, counties, or provinces to be able to function within a European framework.

Among the sedentary indigenous peoples, as in the Iberian system, the household held and worked land and paid taxes. In both, women were in some ways subordinate to men. But in both cultures they could hold and bequeath personal and real property and carry out various kinds of economic transactions, retaining many rights within marriage. In the matter of marriage alliances, crucial to the organization of both kinds of societies, the woman and her property and rank were as important as the man and his.

Semisedentary peoples. Among the semisedentary peoples, much of the above structure was missing. Without well-defined permanent local political units, strong rulers, or tax mechanisms, they did not offer the Europeans the same kind of potential foothold. They lacked social classes, depending on gender and age for their primary social distinctions. Even their household and family structure were different. Settlements or villages shifted over time both in location and in membership; the largest strongly defined unit was a household often containing scores of people related by blood and marriage, headed by the eldest male, and the best-defined duties in the society were internal to the household.

Among the sedentary peoples, men did most of the heavier agricultural work, with help only at times of peak workload from women, who were principally involved in processing and distributing the product, much as in Europe. Among the semisedentary peoples, men mainly hunted, only clearing the fields for the women, who did the bulk of the agricultural work. Warfare was highly developed among both the sedentary and the semisedentary peoples, but the semisedentary were more mobile, were better able to protect themselves in forests and other hazardous environments, and had more effective weapons. Their foods were less attractive to Europeans, and in any case they had less surplus and were fewer in number. They offered Europeans less incentive to invade and more effective resistance when they did.

Warfare

Nonsedentary peoples. With the fully nonsedentary peoples, these factors were multiplied yet again. No agricultural stores at all were available to an invader, nor was there anyone who could readily be compelled to do agricultural work after conquest. The people were extremely few and spread over an enormous territory, able to move long distances at short notice. Their military potential was much greater than that even of the semisedentary peoples. With so little incentive for the Europeans to subdue them, so few points of contact between their societies, and such great ability and will on the part of the nonsedentary peoples to resist conquest, the main patterns between the two groups became avoidance and long-continuing conflict.

THE IBERIANS

In most ways the Spaniards and Portuguese shared the characteristics of other European peoples. They did, however, have some special features as inhabitants of the Mediterranean region and the southwestern part of Europe.

Cities. In the late 15th century most of Iberia was consolidated into three kingdoms—Portugal, Castile, and Aragon—of which the last two were united through royal marriage. But society itself was still quite provincial. The most important entity for purposes of organization and affiliation was the city and the large territory attached to it. More people were engaged in agricultural and pastoral pursuits than anything else, yet society was urban-centred. Each province focused on a city where not only most governmental, ecclesiastical, professional, commercial, and craft personnel congregated but where even the families who controlled the largest rural estates resided. The town council, or cabildo, united representatives of the most prominent families of the whole province, which was thus not divided along urban and rural lines. Rather, a strong solidarity prevailed, with the less successful flowing to the edges, the more successful back to the centre. The cities that the Iberians established in the Americas had the same characteristics, becoming the means of organizing huge territories around a European settlement.

Family. Some characteristics of the Iberian family differed from those found in the northern European family, and these were to have profound effects on relations between Iberians and indigenous people in the Americas. In the Iberian tradition, families were multilinear and existed at different levels. A marriage did not subordinate the wife's family to the extent common in the north of Europe. Women kept their maiden names after marriage, and the dowry given with them remained their own property. Some of the children of a given pair might take the name of one parent, some the name of the other, the choice often being determined by who ranked highest socially. Rather than counting only from father to son to grandson, the Iberians kept track of a network of connections, as many made through the female line as the male.

Marriage and informal unions

Formal marriage was undertaken only when the partners, and especially the male, considered themselves fully established. Men often married quite late, whereas women, for whom the possibilities of advance were severely limited, tended to marry earlier. Many couples never married at all, so that their children were in the strict legal sense illegitimate. While they were waiting, late-marrying men would have relationships with women of lower rank, and children were born of these informal unions. The result was that, despite the ostensible disapproval of the church, Iberian society was full of informal partners and illegitimate children.

A complex set of practices had grown up for the treatment of the women and children involved in informal unions. When the man finally decided to marry, he would often provide for his informal partner, giving her something as a dowry so that she could herself get married to someone of lower rank. The father might recognize the children of these unions, giving them his name and some sort of protection. They were not at the level of his legitimate children, but they were useful as trusted aides or stewards, and he might arrange marriages between the female children and his subordinates. In the Western Hemisphere, the lower-ranking women with whom Iberians had informal unions were often indigenous or African, and the children were racially mixed, but the Iberian patterns of treatment of those involved in the informal unions remained much the same, allowing for a vast amount of social and cultural contact and mixture.

Ethnic diversity and its results. Christians speaking closely related Romance languages made up the majority of the inhabitants of the Iberian peninsula, but they had long coexisted with a larger element of starkly distinct peoples than most of the other nations of Europe. Not only were the Basques in the northeast of different stock, but Iberia had been largely conquered in the early Middle Ages by Muslim Arabic speakers coming from northern Africa across the Strait of Gibraltar. In a long process of reconquest, called the Reconquista, the Iberians had gained back nearly all of the peninsula by the late 15th century, but the Moors, as they called them, were still the majority of the population in several areas along the southern coast, and as servants, slaves, and craftspeople they were to be found in many parts of the peninsula. A substantial number of Jews had also long made Iberia their home. For many decades the Portuguese had been exploring along the coast of Africa, bringing back many Africans as slaves. By the late 15th century Africans were present in considerable numbers in Portugal and also in the south of Spain.

The Iberian Christians' relations with the other peoples, above all the Moors, were to be the precedent for their treatment of the inhabitants of the Western Hemisphere. In the Reconquest (Reconquista) the Christians had pushed their rivals back through military force; those who carried out the conquests often went to settle among the Moors and were rewarded by the government with grants of land and other benefits. But the newly subjugated Muslims retained much of their organization and civilization for long periods, only gradually being Christianized and absorbed. As for the Jews, on the one hand they were resented and sometimes persecuted by Christian Iberians while on the other hand those who converted to Christianity often rose high in professional and political life and married well within Christian Iberian society.

The Reconquest tradition

The Africans had become a well-known group especially in the southern part of the peninsula, with accepted roles as house servants, craftspeople, and field workers. Possession of African slaves was part of general economic life and of social ambitions. Also, manumission was possible, and communities of freed Africans, many of them racially mixed, existed on the edges of society.

So much diversity represented a formidable challenge to the movement toward the creation of unified Christian nation-states that was coming to a head in the late 15th century. Those of the Jews and Moors who had refused to convert were in time forcibly expelled, and the Inquisition became active in the attempt to enforce the orthodoxy of those who had accepted conversion. Negative stereotypes concerning the other ethnicities were rife in Iberian culture, but over the centuries Iberia had seen diversity, close contact with different peoples, and their gradual absorption.

The overseas tradition. All Iberia's coastal peoples had maritime experience. Yet farther inland the occupation of mariner was despised; expansion was deemed a matter of conquering and occupying contiguous territory rather than of far-flung commerce. It was the Italians, above all the Genoese, who brought the lore of overseas activity to the Iberians. From the eastern Mediterranean they carried the sugar industry, the use of foreign slaves in it, and the trinket trade with distant peoples first to Spain and Portugal and then on out into the Atlantic, where they were involved together with the Portuguese on the West African coast and the islands lying off it. By the time of contact with the Americas, the Spaniards had been affected by these developments to the extent that Seville and some other ports were heavily engaged in overseas commerce, often under Genoese direction, but they still mainly adhered to the tradition of conquest and settlement, reinforced by their final defeat of the Spanish Moors in 1492. The Portuguese, on the other hand, partly because of Italian influence and partly because of their own geographic situation, had gone over thoroughly to the commercial-maritime tradition, emphasizing exploration, commerce, tropical crops, and coastal trading posts rather than full-scale occupation.

It is no accident, then, that Christopher Columbus was a Genoese who had long been in Portugal and had visited the Atlantic islands. His projects were entirely within the Italian tradition.

Early Latin America

SPANISH AMERICA

The Spaniards were not only the first of the Europeans to reach the Americas in early modern times, but they also quickly located and occupied the areas of greatest indigenous population and mineral resources. They immigrated in force and created a far-flung, permanent network of new settlements.

The Caribbean phase. The islands of the Caribbean would soon become a backwater, but during the first years of Spanish occupation they were the arena of the development of many practices and structures that would long be central to Spanish-American life.

When Columbus returned to Spain from his voyage of 1492, having hit upon the island of Hispaniola (now divided between the Dominican Republic and Haiti) as his base, his concept of what should be done thereafter was in the Italian-Portuguese maritime tradition. He wanted to explore further for trading partners, and he considered all who came along with him to be employees of an enterprise headed by himself. The Spaniards, however, immediately started moving in the direction of their own traditions. The expedition that returned to Hispaniola in 1493 was far more elaborate than it needed to have been for Columbus' purposes, containing a large number and variety of people, animals, and equipment for a large-scale, permanent occupation of the island. A conflict of purpose between the Spaniards on the one hand and Columbus with his Italian relatives and associates on the other soon ensued. By 1499 the royal government was intervening directly, naming Spaniards to the governorship and sending

Triumph of the Spanish tradition

further large parties of settlers. Spanish ways soon gained the upper hand.

Santo Domingo, founded on the southeastern coast of Hispaniola in 1496, became a real city, with a rash of ephemeral secondary Spanish cities spread over the island. These were oriented to gold-mining sites, which were soon at the base of the Spanish economy. Indigenous demographic loss in this hot, humid area was quick and catastrophic, and the placer mines also soon began to run out. In the second decade of the 16th century the Spaniards pushed on to the other large islands, where the cycle began to repeat itself, although more quickly; around the same time, expeditions to the mainland began, partly to seek for new assets and partly to try to replace the lost population on the islands.

The city. Santo Domingo became a type of entity that would reappear in every major area of Spanish occupation. The central city formed a stable headquarters for the Spaniards in the midst of a chaos of population loss and economic shifts in the countryside. The majority of all the Spaniards in the country lived there, at least when they could. Everyone of importance was there, with only underlings doing essential tasks located in the country. Governmental offices, churches, large private dwellings, and shops soon materialized around the city's central square, together with all the people required for them. The urban core was well laid out and well built up. On the city's edge everything was different. Here were the ranchos, impermanent structures inhabited mainly by Indians temporarily in town for work purposes. The Spanish-American city remained like this for centuries—Spanish in the centre, Indian on the edges, growing indefinitely without changing at the core, the site of an enormous process of cultural change.

Indians and Spaniards. In the Caribbean phase several mechanisms developed, combining indigenous and Spanish elements, that long formed the main structural ties between Indians and Spaniards on the mainland as well. The primary form through which Spaniards attempted to take advantage of the functioning of the indigenous world was what came to be known as the encomienda, a governmental grant of an indigenous sociopolitical unit to an individual Spaniard for him to use in various ways. On the Spanish side, the institution grew out of the Reconquest tradition. Pressure among the Spaniards on the scene led to the arrangement; Columbus, while governor, had opposed it, and Spanish royal authorities tried to restrict it as much as they could. On the indigenous side, the encomienda rested on an already existing unit and the powers of its ruler. The size and benefits of the encomienda thus depended on the local indigenous situation: there could be only as many encomiendas as there were indigenous units; the encomendero (holder of the grant) could at least initially receive only what the ruler had received before him. The larger islands were inhabited by the Arawak, a sedentary if modestly developed people with kingdoms, rulers, nobles, and obligatory labour mechanisms. Their ruler was called a cacique, and the Spaniards adopted the word and carried it with them wherever they went in the Americas. The cacique received labour but not tribute in kind, and the encomendero followed suit.

The encomienda

The encomendero used the indigenous labour in various ways: to construct houses in the Spanish city where he lived, to provide servants, to produce agricultural products on properties he acquired, and above all to work in the growing gold-mining industry. The encomienda set up most of the main forms of Spanish-Indian contact. Although based on traditional mechanisms, it involved major movements of people and new types of activity. Through these dislocations and the exposure of the Indians to new diseases, the encomienda was instrumental in the quick virtual disappearance of the indigenous population on the large islands.

The encomienda was primarily a transaction between the encomendero, the cacique, and his people, but it could not stop there. Auxiliaries with European skills were needed to run mining operations and supervise the growing of European crops and livestock. The encomendero would hire some Spaniards in supervisory capacities, augmented by African slaves when possible, but the limits of his resources were soon reached. He needed permanent indigenous employees who could learn needed skills and act as a cadre. The indigenous world already knew the *naboría,* a person directly and permanently dependent upon the ruler or a noble. This role was appropriated by the Spaniards, who commandeered substantial numbers of Indians for their permanent employ, calling them *naborías.* On the mainland the permanent indigenous worker was to become an ever-growing element of the equation, the locus of the greatest cultural change, and a channel between the Spanish and indigenous worlds.

Indian slaves

In the Reconquest tradition, the Spaniards believed that non-Christians taken in battle could properly be enslaved. Nevertheless, the bulk of the sedentary population in the Caribbean and on the mainland was not enslaved. Only as the population declined seriously did slave-raiding around the edges of the Caribbean become a major factor, the Spaniards attempting in vain to replace the losses. All over Spanish America, Indian slavery was to be a secondary factor, brought into play mainly with less-than-sedentary peoples and under economic pressures—that is, the lack of other assets. The slaves were always, as in this case, employed far from their place and culture of origin.

A new Spanish subculture. *Cacique* was not the only word and concept incorporated into local Spanish culture in the Caribbean and spread from there wherever the Spaniards went. Some of the new cultural goods were the result of Spanish action, like the encomienda or the ranchos; others were straight out of the indigenous world, including *naboría, maíz* (corn; maize), *canoa* (canoe), *coa* (digging stick), and *barbacoa* (grill, palisade, anything with pointed sticks, the origin of the English word barbecue). Still others came out of the Portuguese Atlantic tradition, like *rescate* (literally rescue or redemption), a word for informal trading with indigenous people often involving force and taking place in a setting where conquest had not yet taken place. This whole new overlay on Hispanic culture maintained itself partly because it was adjusted to the new situation but above all because each set of new arrivals from Spain readily adopted it from the old hands already there.

Conquest in the central mainland areas. The Spanish occupation of the larger Caribbean islands did not entail spectacular episodes of military conflict. Yet force was involved, and the Spaniards developed many of the techniques they would use on the mainland. One of the most important was the device of seizing the cacique in a parley, then using his authority as the entering wedge. The Spaniards also learned that the indigenous people were not a solid unit but would often cooperate with the intruders in order to gain advantage against a local enemy.

Also during the Caribbean phase an expeditionary form evolved that was to carry the Spaniards to the far reaches of the hemisphere. Spanish expansion occurred under royal auspices, but expeditions were conceived, financed, manned, and organized locally. The leaders, who invested most, were senior people with local wealth and a following; the ordinary members were men without encomiendas, often recently arrived. The primary leader of an important expedition was often the second-ranking man in the base area, just behind the governor, ambitious to be governor himself but blocked by the incumbent.

There was no permanent organization and no sense of rank. The word army was hardly used, and the word soldier not at all; still, the possession of steel helmets, steel swords and lances, and horses gave the Spaniards an overwhelming technical advantage over any indigenous force they were likely to meet. On flat, open ground, two or three hundred Spaniards often defeated indigenous armies of many thousands, suffering few casualties themselves. The conquering groups showed a surprising diversity, coming from many different regions of Spain (plus some foreign countries) and representing a broad cross-section of Spanish pursuits. It was they who founded and settled in the new cities, and the later stream of immigration initially consisted primarily of their relatives and compatriots. Conquest and settlement were a single process.

Having in about one generation largely exhausted the demographic and mineral potential of the greater Antilles, the Spaniards began a serious push toward the mainland in two approximately contemporary streams, one from Cuba to central Mexico and surrounding regions and the other from Hispaniola to the Isthmus of Panama region and on to Peru and associated areas. The Peruvian thrust started first, in Tierra Firme (the area of Panama and present northwestern Colombia) in the years 1509–13. The results were appreciable, but the Panamian occupation was thrown somewhat in the shadow for a time by the spectacular conquest of central Mexico in 1519–21.

Conquest of Mexico. The leader of the Mexican venture, Hernán (Hernando) Cortés, had some university education and was unusually articulate, but he conformed to the general type of the leader, being senior, wealthy, and powerful in Cuba, and the expedition he organized was also of the usual type. Passing by the Maya of the Yucatán Peninsula, the Spaniards landed in force on the central coast, almost immediately founding Veracruz, which despite small shifts in location has been the country's main port ever since. The Aztec empire, or Triple Alliance, of the city-states of Tenochtitlán, Texcoco, and Tacuba, centring on the Mexica (Aztec) of Tenochtitlán, dominated central Mexico. The coastal peoples among whom the Spaniards landed, however, had only recently been incorporated in the Aztec tribute system, and they offered the Spaniards no open resistance.

Cortés' expedition

Moving inland, the invaders encountered the second power of the region, the Tlaxcalans. Tlaxcala briefly engaged the Spaniards in battle but, suffering heavy losses, soon decided to ally with them against their traditional enemy, the Aztec. As the Spaniards moved on toward Tenochtitlán, many of the local subordinate states (*altepetl*) also came to terms. Even in Tenochtitlán itself fighting did not ensue immediately; the Spaniards as usual seized the cacique (that is, the king of Tenochtitlán, often called the Aztec emperor, Montezuma or Moteucçoma) and began to exercise authority through him.

The expected secondary reaction was not long in coming, and fighting broke out in the capital. At this point the most unusual part of the process began, for Tenochtitlán was on an island in the midst of a lake, shot through with canals and extensively built up. Here the Spaniards lost much of their usual advantage. They were forced from Tenochtitlán with severe casualties. Although they retained their superiority in the open country, they had to retire to Tlaxcala, accumulate reinforcements, and then come back to Tenochtitlán to carry out a unique full-scale siege, including the use of European-style vessels with cannon on the lake. After four months the Spaniards captured the Aztec capital and began turning it into their own headquarters as Mexico City.

Other parts of central Mexico came under Spanish control more easily, and several Spanish cities were established in the region. Soon successor conquests were under way, to Guatemala, Yucatán, and the north. Those to the north led to little in the short run because that area was inhabited by less-sedentary peoples. Cortés acted as governor for a time and was given great rewards, but rivalries among the Spaniards soon made it possible for the royal government to replace him, first with an audiencia, or high court, and then also with a viceroy, direct representative of the Spanish king.

Conquest of Peru. The Spanish thrust toward Peru through Panama was diverted for some years by the attractions of nearby Nicaragua. No one knew what lay along the southern coast, which because of contrary winds was very difficult to navigate; the coastal climate was hostile, and little wealth was discovered among the people dwelling there. Attempts in this direction were led by Francisco Pizarro, who despite being illegitimate and illiterate had all the other familiar characteristics of the leader; not only was he the illegitimate son of a prominent family but he also was one of the first captains on the American mainland, by the 1520s a wealthy encomendero and town council member of Panama. At length Pizarro's group came into contact with central Andean coastal people connected with the Inca and saw evidence of great wealth and development. Acquiring from the crown the governorship of the new region, which now began to be called Peru, Pizarro, in 1530, led an expedition that proceeded into Inca territory. In 1532, at the north-central site of Cajamarca, the Inca emperor Atahuallpa was captured in the usual fashion, a parley and surprise attack. In 1533, after much treasure had been collected, the Spaniards had Atahuallpa executed.

The process of conquest and occupation was much as in Mexico, though Pizarro was not thinking of Mexican precedents. Again, once the Spaniards were in the fully sedentary lands of the Inca, the local people hardly attacked them, allowing them to proceed unhindered into the very presence of the imperial ruler. In addition to a localism similar to that of Mexico, the situation was defined by a large-scale Inca civil war that was just ending as the Spaniards arrived. A faction based in Quito, headed by Atahuallpa, had defeated a faction based in Cuzco, the traditional Inca capital, but the victory had not been entirely consummated, and the parties were still very bitter. After the events at Cajamarca, the Spaniards faced a certain amount of fighting as they advanced to Cuzco, especially from adherents of Atahuallpa, but his enemies, who seem to have been the majority on the ground, tended to acquiesce for the time being.

The Spaniards founded a major Spanish city in Cuzco, but they stopped short of making it their capital as their compatriots had Tenochtitlán in Mexico. Deterred by the rigours and inaccessibility of the southern Peruvian highlands, after a bit of experimentation they established the new settlement of Lima, on the central coast, as capital of Peru. The move was of vast significance. In Mexico the bulk of the Spanish population concentrated in the area of highest indigenous population density, favouring contact, cultural change, and merging. In Peru, the highland centre of indigenous population was separate from the centre of Spanish population on the coast, which, in addition, quickly lost most of its indigenous inhabitants to disease. In consequence, the two peoples and cultures underwent an overall slower and less thorough process of amalgamation.

Lima

As in Mexico, conquering expeditions soon went out from central Peru, in all directions: to Quito and on north to Colombia, to Chile and Argentina to the south, and even to the Amazon. Peru proper seemed to be securely conquered, but a countrywide uprising took place in 1536, centring in Cuzco, where the Spaniards were kept surrounded for more than one year, until an expedition returning from Chile lifted the siege. After that, the conquest was definitive, although the successor to the Inca ruler and a group of followers took refuge in a remote region, where they held out for more than a generation.

Peru's history continued to be less placid than that of Mexico. Peru was much harder to reach from Spain, and travel within the country was extremely difficult. In the conquest period and long after, Peru was far richer in precious metals than Mexico, since the Spaniards profited from the silver mining already developed by the Inca. Thus there was more to fight over, and struggles arose between the Pizarro brothers (Francisco had three) and a faction led by Diego de Almagro, Pizarro's junior partner. Spaniards flooded into the country, eager for encomiendas and ready to rebel in order to get them. Four large-scale civil wars among the Spaniards rocked the country in the time between the late 1530s and early 1550s.

Like Cortés and like most leaders of successful expeditions, Pizarro became governor of the country he had conquered and actually held that position longer than Cortés. In 1541, however, he was assassinated, brought low by the second of the Almagrist rebellions. A royally appointed governor from outside took over, followed in 1544 by a viceroy and audiencia based in Lima; the first viceroy was in turn killed in a civil conflict, but his successors became more firmly established.

Conquest society in the central mainland areas. In the generation or two subsequent to the military phase of the conquest, Spanish immigrants poured by the thousands into Mexico and Peru. Although still a small minority compared with the indigenous population, they constituted the great majority of all Europeans in the hemisphere, so that these two regions could now be doubly called central areas. They combined the largest European and indigenous populations with the liveliest economies, for they proved to be the sites of the richest deposits of precious metals then known. The immigrants continued to come from all parts of Spain, constituting an even broader cross section than had the conquerors, for women were now a standard part of the stream.

The central-area encomienda. Already crucial in the Caribbean, the encomienda now developed even further. The Mexican and Andean indigenous units on which it was based were much larger, with stronger authorities who could collect tribute in kind as well as labour. Moreover, the products could circulate in an economy with a great deal more liquid wealth, and there were now many more non-encomenderos, who soon formed the great majority of all Spaniards. The encomenderos greatly enlarged their staffs and followings, with various levels of stewards and many more African slaves, whom they could now afford. The ecclesiastics who now began serious work with the indigenous people of the countryside operated within the framework of the encomienda and received their remuneration from it. The encomenderos went not only into mining and local agrarian activity on a larger scale than before but also into a large variety of ancillary enterprises. Their establishments in the city centre were often palatial, including shops rented to merchants and artisans, of whom they were the best customers. They married Spanish women, ideally relatives of other encomenderos or of high local officials, if only to have legitimate heirs to inherit the encomienda. They became an interlocking group dominating local Hispanic society and virtually monopolizing the municipal councils of the Spanish cities. The process whereby Hispanic society penetrated into the hinterland was begun by their usually humble rural employees, often called *estancieros,* who combined tax collecting, labour supervision, farming, and livestock growing.

Artisans. The Spanish crafts flourished in the encomenderos' cities, practiced by artisans who had a far humbler social profile than the encomenderos but were like them in being tied to the locality. They, too, frequently married Spanish women and acquired urban and rural property. To increase their productivity, they bought African slaves, whom they trained in their own trades; the Africans in turn helped train the larger number of Indian apprentices to be found in many shops. In this way the artisans were important in the gradual creation of an ever-growing African, indigenous, and mixed group in the cities, able to speak Spanish and practice the Spanish trades.

The role of Africans

Spanish women. Spanish women were an important element in the sedentary urban society growing up in the central areas. The women were above all relatives of Spanish men already present, brought from Spain explicitly to marry some local associate. As wives of encomenderos and artisans, they managed households that included many Spanish guests and employees and even larger numbers of Africans and Indians, whom they attempted to mold to their purposes. They also brought up both their own fully Spanish children and the racially mixed children they often took or were given to raise. As widows and sometimes spinsters, they actively participated in economic life, though women's independent activity tended to be channeled into certain conventional directions, from indirect investment and owning urban real estate at the higher levels to running bakeries and taverns at the lower. Women were at first a small minority of the Spanish population, but their relative numbers steadily increased, reaching effective parity with men by the second or third generation after conquest.

Africans. Africans also were important to the society. As stated, encomenderos and artisans acquired African slaves, and any Spaniard of means would try to own at least one or two. Thus Africans were soon a significant group numerically; on the Peruvian coast, at least, it is thought that after several decades they equaled the Spaniards in numbers. Spaniards needed auxiliaries serving as intermediaries between themselves and the much larger indigenous population. Africans, who shared the Spaniards' Old World immunities and much else, survived and adapted well; the main limitation on acquiring them was the great expense involved.

The gender ratio strongly favoured males, but females were present too, usually in household service, food trades, and petty commerce. The women were frequently mistresses of their owners, to whom they bore mulatto children, with the result that mother and children were sometimes freed. Other African slaves bought their freedom, and a mainly urban class of free blacks began to emerge. Their roles were similar to those of the slaves, except for being exercised more independently.

In this society, the slave, and especially the African slave, was not at the bottom of society but ranked in Spanish terms higher than the general Indian population. Africans were more closely associated with the Spaniards than Indians, culturally more like them, given more skilled and responsible tasks, and in cross-ethnic hierarchies were normally in charge of indigenous people.

Indians among Spaniards. Spanish cities, from the very beginning, were full of Indians working for Spaniards in a great number of capacities, sometimes temporarily, sometimes for long periods, but usually at a low level. One of the most important features of life in the first postconquest decades was the prevalence of Indian servant-mistresses of Spaniards, the result of the fact that Spanish women were still much less numerous than men, not to speak of the pattern of men waiting for full success before marrying. These indigenous women retained many aspects of their traditional culture, but they had to learn good Spanish and master skills of Spanish home and family life. They bore the Spaniards mestizo children, who were to become a very important feature of postconquest society.

Indian servant-mistresses

Commerce. Merchants were present in force and vital to the existence of the overall complex. But as members of a far-flung network that required high geographic mobility, they were at first less a part of local society. Once the wealth of the central areas became apparent, Seville-based firms began to dominate the import-export trade—the exchange of American precious metals for European cloth, iron, manufactures, and other goods. The representatives at American ports and capitals were junior partners in transatlantic firms and in time expected to move on; hence they seldom married or bought property locally. The aim was to get silver back to Seville in order to pay debts and reinvest in merchandise. Second-rank merchants, however, without direct ties to Seville, were more likely to develop local roots.

Commerce in local goods, often but not always of indigenous origin, was carried on by members of a well-defined social type, sometimes called *tratantes,* with a profile sharply distinct from that of the long-distance merchants. Often illiterate, and furthermore without capital, they were recruited from among the most marginal members of local Hispanic society. They, too, were relatively unstable; they were prone to move to another area or into other kinds of activity because their status was so precarious.

Mining. The mining sector drove the economy of the Spanish world and was an indispensable component of it, yet in several ways it stood apart. It employed only a relatively small proportion of the total Spanish population. Mining complexes were mostly remote from the main centres of indigenous settlement and hence also from the network of Spanish cities. Turnover was quick, whether in terms of sites, mining enterprises, or individuals.

Gold mining was often virtually an expeditionary activity; a gang of Indians, joined perhaps by some blacks and led by one or two Spanish miners, might spend only days or weeks at a given river site. An encomendero, not himself physically involved, would likely supply the finances and take most of the profit. In many regions gold mining was seasonal, with miners having neither special training nor a full commitment to the industry.

In most regions placer gold was soon exhausted, though Mexico was forced to rely on it for a generation. Silver mining was the successor, and it became the main export asset of the central areas until the time of independence. Here too the encomenderos were the greatest investors and mine owners in the beginning, but their dominance was short-lived. Silver mining was the type of technically demanding, capital-intensive enterprise that called for close attention and much expertise on the part of owners. Very soon true silver mining experts began not only to operate the mines but to become the owners as well.

Spanish law granted the crown residual ownership of mineral deposits, giving it the right to levy substantial taxes on the industry. There was always a governmental presence at mining sites, and the silver tax was the crown's principal source of revenue. Silver mining camps began to resemble ordinary Spanish municipalities, with councils (dominated by local mining entrepreneurs) and strong contingents of merchants, craftspeople, and professionals.

Differences between Mexican and Peruvian silver mining industries

By 1550 strong differences had developed between the Mexican and the Peruvian silver mining industries. In the Andes the great deposits, of which those of Potosí Mountain (in present Bolivia) were overwhelmingly predominant, were within the territory of sedentary indigenous population; moreover, the Andeans had a strong tradition of long-distance labour movements. Thus indigenous labour obligations, channeled first through the encomienda and later through other arrangements, could supply a large stream of temporary workers. In addition, there were a number of permanent indigenous workers, some of whom possessed skills inherited from the preconquest period, and, in an industry as technical as mining, this group was constantly growing. Even so, the Peruvian mines used large numbers of temporary labourers under governmental obligation, and their presence greatly slowed down cultural change among the indigenous mine workers.

In Mexico, most of the largest silver mining sites were discovered well to the north of the zone of sedentary population. Traditional labour obligations could not be used, and the bulk of the labour force consisted from the beginning of sedentary Indians from the centre acting as free agents—*naborías,* or permanent workers. The Mexican mines also used far fewer people, so that the Hispanic element predominated more than in Peru, and the north of Mexico was soon on its way to having a Hispanized, mobile population very different from that in the central part of the country.

Institutional, legal, and intellectual developments. From early in the Caribbean phase the crown had established the Casa de Contratación, or board of trade, in Seville, apparently originally intended to operate the entire overseas enterprise on an Italian model. In fact, it soon became a customs and emigration office, involved also in the organization of Atlantic convoys. Direction of the governmental aspect of overseas life went to a royal council constituted much like others, the Council of the Indies (as the Spaniards continued to call America), which issued decrees, heard appeals, and above all made appointments to high offices. Distances were such that almost everything governmental depended on the officials actually in America.

During the conquest and immediately thereafter, royal government was nominal in the sense that the governor was invariably merely the leader of the conquering expedition. But in the central areas, with the rivalries and wars among the conquerors and continued strong Spanish immigration, the royal government was soon able to install its own institutional network, with the support of many local Spaniards. As stated earlier, before 1550 both Mexico and Peru had a viceroy and an audiencia, based in the respective capitals, and some secondary audiencias followed; there were substantial treasury offices as well, for the crown's most urgent interest in the new areas was getting silver revenue. A host of lawyers and notaries assembled in the capitals around these nuclei and their branches in the secondary Spanish cities. The viceroys brought with them retinues including an element of high nobility. Marriage alliances and business deals soon brought the officials into connection with the more important encomenderos.

Church organizations

Church organizations, which in the Spanish scheme of things were part of the overall governmental framework (the crown appointed bishops and many other high officials of the church), also came into the central areas in force on the heels of the conquest. Few clerics of any kind were with the actual conquering expeditions, but soon parties of friars arrived. They were followed by bishops and cathedral chapters, established first in the capitals and then in secondary cities; the culmination of the process was the seating of archbishops in Lima and Mexico City. Both the friars and the priests began to penetrate the countryside, operating through the encomiendas, with the ideal (long unrealized) of having one cleric for each encomienda. Like the governmental officials, ecclesiastics were closely connected with the civil society; some were appointed in the first place because of family connections, and many tried to marry female relatives to encomenderos.

These institutions were an important part of the general scheme, but they depended on local Hispanic civil society and reflected its relative strength or weakness. Governmental and ecclesiastical hierarchies were as urban-oriented as all other aspects of Spanish society; they were based in the cities, above all the largest cities, where one could find not only the largest concentrations of personnel but all those of high rank. The religious orders were a partial exception, rotating their members frequently; nevertheless, the most famous figures spent the bulk of their lives in larger centres. As for the government, it hardly existed outside the cities; the local magistrates who gradually came to be appointed in the Indian areas were mainly laymen, often unsuccessful candidates for encomiendas.

In the aftermath of the conquests, as they became integrated into the local situation, some ecclesiastics began to criticize Spanish institutions, especially the encomienda. However, the various representatives of the church were not entirely unified. The secular clergy said little; among the orders, the pragmatic Franciscans wanted a higher moral tone and better treatment of the Indians but were prepared to work through the encomienda; the more doctrinaire Dominicans, of whom Bartolomé de las Casas was the most famous and most persistent, spoke for the total abolition of the encomienda, with the clergy to be in charge of the Indians. At the same time, the Spanish royal government was seeking to find ways to increase its authority and in alliance with the Dominicans passed antiencomienda legislation. Resistance among the settlers and conquerors was fierce (the greatest of the Peruvian civil wars was in direct reaction to the strongest legislation, the New Laws of 1542). But in combination with other factors (of which indigenous population loss and the presence in the central areas of many non-encomenderos were the most essential), in the course of the 16th century the encomienda lost its labour monopoly and had its tribute in kind curtailed, while many encomiendas without legal successors reverted to direct crown administration.

The conquerors and early settlers produced a large number of histories describing and praising their exploits. The ecclesiastics, as they came in, began to write similar documents about their own activities, but they also went much further. Some, with the Franciscans most prominent, showed a strong interest in the study of indigenous history, language, and culture; others, especially the Dominicans, wrote in a more polemical spirit; and sometimes the two currents converged. The arts of literacy were much prized by the upper levels of the Spanish population, and universities, mainly for professional training, were soon established in the viceregal capitals.

Trunk lines. Not only were the central areas different from the fringes in early Latin America, but important distinctions existed within the central areas themselves. In some ways the centre was more a line than a region—that is, a line from Atlantic port to capital to mines, along which European people and products flowed in and silver flowed out. For Mexico, the line went from Veracruz to Mexico City and on to Zacatecas and other mines of the north. In the more complex Peruvian scheme, the line went from the Isthmus of Panama to Lima and on to Potosí. It was along these routes that the Spanish and African populations concentrated, that social, economic, and governmental institutions were first created, then gelled and thickened, and that cultural and social change proceeded most quickly.

Postconquest indigenous society. Although the majority of the indigenous population continued to live in their traditional units across the countryside, their lives were nonetheless profoundly affected by the conquest and its aftermath. The most obvious development was drastic demographic loss; in a process marked by periodic large epidemics, the population declined through the 16th century and on into the 17th century to a small fraction (impossible to determine with precision) of its precontact size. Only in hot, low-lying areas, such as the Peruvian and Mexican coastal regions, however, were losses as disastrous as those of the Caribbean islands. The peoples of the temperate highlands, however much they may have diminished in numbers, survived in the sense of retaining their local units, their language, much of their cultural heritage, and the essence of their social organization.

Population loss

The Nahuas of central Mexico are the people whose postconquest experience is best understood because of the voluminous records they produced in their own language. These records reveal that the Nahuas were not overly concerned with the Spaniards or the conquest, which seemed to them at first much like earlier conquests; they remained preoccupied to a large extent with their internal rivalries. The local state, the *altepetl,* with its rotating constituent parts, remained viable as a functioning autonomous unit and as bearer of all major Spanish structural innovations, not only the encomienda but also the parish and the indigenous municipality. The Nahuas accepted Christianity and built large churches for themselves, but those churches had the same function as preconquest temples, acting as the symbolic centre of the *altepetl,* and the saints installed in them had the same function as preconquest ethnic gods. The status and duties of the commoners remained distinct from those of the nobles, who manned the local Hispanic-style government of the *altepetl* as they had filled offices in preconquest times.

The household and land regime remained much the same in its organization despite reductions and losses. Household complexes, for example, continued to be divided into separate dwellings for the constituent nuclear families. The Spanish concept "family" had no equivalent in Nahuatl, and none was ever borrowed. The greatest internal social change was a result of the end of warfare, which had been endemic in preconquest times. Performance in war had provided degrees of social differentiation, avenues of mobility, and a large supply of slaves. Formal slavery among Indians soon disappeared, while internal social mobility tended to take the form of commoners claiming to be nobles or denying specific rights to specific lords. However, the categories themselves were not challenged: the strong distinction between commoner and noble was not soon erased. An entirely new type of mobility had come into existence—movement of Indians away from the whole realm of indigenous society in the direction of the Spanish world to become *naborías* or city dwellers.

The peoples from central Mexico to Guatemala had had forms of recordkeeping on paper in preconquest times, and after the arrival of the Spaniards a remarkable cooperation between Spanish ecclesiastics and indigenous aides led to the adaptation of the Latin alphabet to indigenous languages and subsequently to regular record production. In the case of Nahuatl, the main language of central Mexico, the records have allowed the tracing of some basic lines of cultural and linguistic evolution in three stages. During the first generation, although cataclysmic change was occurring, Nahua concepts changed very little, and their language could hardly be said to have changed at all, using its own resources to describe anything new. In a second stage, beginning about 1540 or 1545 and lasting for nearly 100 years, Nahuatl borrowed many hundreds of Spanish words, each representing a cultural loan as well. But all were grammatically nouns; other innovations in the language were minimal. This was a time of change in a familiar corporate framework, centring on areas of close convergence between the two cultures. A third stage began about the middle of the 17th century, when Spaniards and Nahuas had come into closer contact, and many Nahuas were bilingual. Now there were no limitations on the kinds of things introduced into the language, and change increasingly took place at the level of the individual, with mediation no longer necessary.

Three stages

The Nahuas had structures perhaps more similar to those of the Spaniards than any other indigenous group, and nowhere else was there such massive interaction of Spanish and indigenous populations, but broadly similar processes were at work across the central areas. Among the Maya of Yucatán, the direction and nature of the evolution was closely similar but much slower, corresponding to the relatively small Spanish presence there. The Yucatec Maya language stayed in something comparable to the second stage of Nahuatl for the entire time up to independence.

In the Andes too the indigenous social configuration was sufficiently close to the Spanish that it could serve as the basis for institutions such as the encomienda and parish. But Andean sociopolitical units were less contiguous territorially than those of central Mexico or Spain, and the population engaged in more seasonal migration. Thus the local ethnic states of the Andes, comparable to the *altepetl* of the Nahuas (though far less well understood) as the framework of social continuity, may have come under greater challenge of their essential character and identity. The Spaniards tended to reassign noncontiguous parts of one entity to other entities geographically closer, thereby mutilating the original entity. As far back as can be traced, the postconquest Andeans were inclined to migrate permanently from their home entity to another, whether to avoid taxes and labour duties or for other reasons. Such movement occurred in Mexico too, but there the new arrivals tended to melt into the existing entity, whereas in the Andes they remained a large separate group without local land rights or tribute duties, known in Spanish as *forasteros*. Another challenge to indigenous society came in the later 16th century in the form of attempts by the Spanish government to reorganize sociopolitical units, nucleating the population in so-called *reducciones,* with consequent social upheaval. Still another apparent disruptive force was the Spanish use of obligatory rotary labour of large groups for relatively long periods at great distances. Yet given the mobility of the Andean peoples from preconquest times, strong continuities may have been involved.

The Andeans had sophisticated recordkeeping systems in preconquest times but did not put records on paper with ink, and after the conquest they did not engage in alphabetic writing on the same scale as the indigenous people of Mesoamerica. Some indigenous-language records are now beginning to come to light, however, and so far cultural-linguistic evolution appears far more similar to that of central Mexico in nature, staging, and timing than one would have expected.

The central areas in the mature period. In the 1570s and '80s the central areas went through a process of codification and institutionalization marking the beginning of a long time of slow transformation, which can be called the mature period. Among the new institutions were those formalizing functions that had long been evolving, including the *consulados,* or merchant guilds, of Mexico City and Lima and tribunals of the Inquisition in the same places (plus Cartagena on the Colombian coast). Entirely new was the Jesuit order, which entered in force at the beginning of this time, quickly becoming strong in urban areas. During these decades nunneries inhabited by daughters of substantial Spanish families came to be a normal feature in any thriving city.

The Jesuit order

Intellectual production began to include not only narrow chronicles but also broad surveys of the entire Spanish-American scene, whether religious, legal, or general in focus. For a time most of the writers were acquainted with both hemispheres, but by the later 17th century locally born Spanish figures were becoming prominent, such as the famous poet, dramatist, and essayist Sor Juana Inés de la Cruz, a Jeronymite nun of Mexico. The late 16th and early 17th centuries saw much significant writing by indigenous authors, affected by both Spanish and indigenous traditions. A large corpus appeared in the Nahuatl language of central Mexico. In Peru the indigenous historian and social commentator (don) Felipe Guaman Poma de Ayala produced a vast work in Spanish.

An elaborate ecclesiastical art and architecture flourished in the main centres, much of it with a special regional style of its own. Religious devotion became more localized, with the appearance of locally born saints and near-saints, notably St. Rose of Lima (Santa Rosa de Lima), as well as miraculous shrines, of which the most famous came to be that of the Virgin of Guadalupe near Mexico City.

Cities and satellites

The Hispanic sector continued to grow, still centred in the same cities founded in the conquest period. These cities maintained their dominance because they attracted to them anyone from the countryside who was fully successful in any endeavour. They were usually filled to overflowing, and consequently they ejected large numbers of lower-ranking Hispanics into the surrounding countryside. As a result, new nuclei of Spanish society began to form outside the cities. The process of urban formation repeated itself; a new entity came into being, Spanish at the centre, Indian at the edges, very much a replica of the original city, except that none of the Hispanics rose above a certain rank, and the whole settlement remained dependent on its parent. In time, under the right conditions, tertiary Hispanic-Indian satellites would arise around the secondary centres in turn, until the entire area was honeycombed, and the original pattern of Spanish city and Indian countryside was obscured.

Racial and cultural mixture complicated and blurred society greatly after the conquest period, but many social criteria were still the same under the surface. The intermediary functions were still the province of those ranking lowest in Hispanic society, but that stratum now contained not only the least senior members (new immigrants from Spain and other European countries) and Africans but also large numbers of mestizos as well as mulattoes and increasingly even Indians who had mastered Spanish language and culture. To organize the diversity, the Spaniards resorted to an ethnic hierarchy, ranking each mixed type according to its physical and cultural closeness to a Spanish ideal. As mixture proceeded across the generations, the types proliferated until finally, at the time of independence, the system collapsed under its own weight. The new categorizations were all at the intermediary level; despite them, all these people, often simply called *castas,* assimilated to each other and intermingled, occupying the lower edge of Hispanic society. The more successful and better connected among them were constantly being recognized as Spaniards, as a result of which the Spanish category grew far beyond simple biological increase and included many people with some recognizably non-European physical traits.

Silver mining in Peru and Mexico continued along the same lines as before, reaching new heights of production in the early 17th century. After that a series of problems reversed the trend for a time. The absolute value of transatlantic trade seems to have fallen during the same period. Scholarly controversies about the existence, nature, and extent of a general economic depression during the 17th century have not been entirely resolved, but it is certain that the expansion of the Hispanic sector of society did not halt.

The most profitable mercantile operations still involved the trade of silver for European products, but some structural changes were occurring. Most of the transatlantic firms of the conquest period had broken up by now. The merchants in the large Spanish-American centres were still mainly born in Spain, but, rather than being members of Spanish firms, they were likely to be agents working on a commission basis or to be operating independently, buying up goods from Spain that arrived in the annual fleets. The change of company structure brought with it a localization of the merchant corps, who now stayed permanently in America, married locally, bought property, and even acted as governmental officials, especially in the treasury and the mint.

This time saw the rise of forms of economic activity not present or not well developed in the conquest period, of which haciendas (landed estates) and *obrajes* (textile shops) are the most prominent. The social organization of such enterprises, however, was familiar from earlier encomienda operations, consisting of a city-dwelling owner, often somewhat removed from daily operations; one or more majordomos; foremen; skilled permanent workers (functional descendants of the *naborías*); and less-skilled temporary workers. The owner was usually Spanish, the middle levels poorer Spaniards or *castas,* and the temporary workers generally still Indians. A powerful trend, corresponding to the growth of city markets and ethnic-cultural changes, was an increase in the proportion of personnel in the middle levels and a decrease in those at the lowest, especially an increase in permanent workers at the expense of temporary ones (though the latter were still very numerous).

Changes in indigenous society

All these developments ultimately had an immense effect on society in the indigenous entities of the countryside. In time, many rural Indians were absorbed within Hispanic society, while leading members of local indigenous society would ally and even intermarry with the humble Hispanics who were now beginning to dominate the local economy. Ties to specific local Spaniards and Spanish organizations gained ever greater importance in the lives of the indigenous people, compared with their own corporate society; one result was large-scale fragmentation of indigenous entities. In central Mexico, many *altepetl* broke into their constituent parts, and in the Andes even many of these constituent parts (*ayllus*) went out of existence or changed their principles of organization.

The Spanish fringe. From the notion of "centre" as used above it follows that the remaining area of Spanish occupation was, from the Spanish point of view at least, peripheral. Most of the Hispanic territories in the Indies were occupied by groups coming precisely from the central areas. Conquering groups had always consisted largely of people of lesser position in the base area, and, as it grew clearer that the central areas were unequaled in their assets, the marginality of the personnel going elsewhere became even more pronounced. In addition to being new and uprooted, those who went to places like Chile, Tucumán (northwestern Argentina), or New Granada (Colombia) were likely to be estancieros and *tratantes* in the centre—not well-born, well-educated, or well-connected. Among them were a larger than average share of non-Spanish Europeans and free blacks. Since these movements were posterior to the initial conquests, the first Hispanics arriving often included some mulattoes and mestizos born in the centre.

Even so, the first Spanish groups in the peripheral areas were comparable to the first conquerors of the central areas in being of varied origins and commanding a variety of necessary skills. A greater difference showed itself later. The central-area conquerors, having struck it rich, sent out appeals to Spain that attracted huge numbers of people, especially male and female relatives, as well as fellow townspeople and others. Fringe-area conquerors had not struck it rich. They were less able to pay for the passage of relatives and less able to attract people in general. As a result, subsequent immigration to the periphery was a much thinner stream than to the centre, and sometimes nearly nonexistent for long periods of time, as in Paraguay, and many activities that were profitable in the centre were not viable. Hispanic society on the fringe was characterized, then, by its relatively small size and slow growth and by weakness in or a total lack of a series of characteristic signs of the centre indicating vigorous development: Spanish women; practicing Spanish artisans; transatlantic merchants; and Africans. The institutional overlay was a mere shadow of the complex network of the centre. The silver-mining sector was entirely absent, though some areas maintained gold production as a second-best (Chile for a substantial period and New Granada indefinitely and on quite a large scale).

From the above it is clear that society on the fringe was less differentiated than in the centre. Also, the encomenderos never rose very far above the rest. Here, the indigenous people hardly knew tribute, and their labour could not be turned into large revenue; moreover, there were far fewer of them. More Spanish intervention was needed, and yet there were not many Spaniards available. Encomenderos on the fringe usually lacked a large staff of majordomos and estancieros. Since the Indians of these regions were organized in much smaller units than those of the centre, many more encomiendas had to be granted among a much smaller number of Spaniards, so that the proportion of encomenderos was greater. Encomenderos and others had to fulfill several functions simultaneously.

When any of these societies began to prosper, however, sharper categorization reappeared, along with a general approximation of central-area patterns. Areas that in one way or another were equipped to supply regions on the trunk line (Guatemala, Venezuela, Chile, and northwestern Argentina) moved most quickly in that direction.

On the fringes, even in regions where it proved possible to establish some form of the encomienda, the relationship between Hispanic and indigenous societies was not the same as in the centre. In extreme cases, as in Paraguay, one can hardly speak of two separate worlds at all; there, in order to take advantage of the largest effective structure the indigenous people possessed—the extended household—the Spaniards actually entered into those households as heads. This led to a permanent indigenous influence on Spanish Paraguayan family structure, customs, diet, and language in a way and on a scale without parallel in the centre. Something of the same effect is observable even in situations where indigenous society was somewhat more like that of the centre, as in the central valley of Chile. The Spaniards dealt with the Indians directly, in small groups or as individuals, so that the distinction between encomienda Indians and *naborías,* so clear in the centre, hardly existed after a time.

Another effect of the nature of the more diffuse indigenous society was that in fringe areas the city, which in the centre was the stable bulwark of Hispanic society, was often notably unstable, shifting from one site to another because no location was predetermined by indigenous settlement. Similarly, rural church activity in the central areas was built squarely on existing territorial and sociopolitical units, using indigenous organization and customs. On the fringe the church for Indians, which here can be called a mission, was founded on a site more arbitrarily chosen, to which indigenous people were attracted, changing their settlement pattern and way of life. The late-arriving Jesuits, who had missed out in the ecclesiastical occupation of the hinterland in the central areas, took a large part in this movement, with especially prominent theatres of activity in the north of Mexico and in Paraguay. The fringe also saw of necessity the building of forts and the creation of standing military forces, paid, if poorly, by the royal government.

Interpenetration of the two societies occurred mainly when the Indians were semisedentary; where they were truly nonsedentary, another pattern emerged. Here the relationship between Spaniards and Indians was of long-standing hostility, with a minimum of social intercourse. Indigenous society remained quite radically separate from Hispanic as long as it survived, whereas the local Spanish societies, though often little developed, were more purely European than in any other kind of situation; the only indigenous people there were usually uprooted sedentary Indians from neighbouring regions. The far north of Mexico and the far south of Chile are two such areas.

In general, one notes a slow tempo on the fringe, with the result that eventually many forms on the periphery seem archaic. The fringe areas tended to maintain some form of the encomienda far into the 18th century, when it was forgotten in the centre; likewise Indian slavery, as well as parish activity among Indians by members of the religious orders, persisted indefinitely. The use of titles was conservative, and many of the social complexities evolving in the centre were slow to reach the periphery.

BRAZIL

The Treaty of Tordesillas (1494) between Spain and Portugal, dividing the non-European world between them, gave the Portuguese a legal claim to a large part of the area to be called Brazil. The Portuguese came upon the Brazilian coast in 1500 on the way to India and would doubtless have acted much as they did with or without the treaty. For decades Brazil was doubly a fringe area. In the Portuguese scheme, it was far behind longer-established and more profitable overseas ventures in Africa and India. In the context of the Western Hemisphere, it was an area lacking known large deposits of precious metals and possessing a semisedentary Tupian population similar and related to the Guaraní the Spaniards were to find in Paraguay; thus it had much in common with the Spanish-American periphery.

The early period. The Portuguese at first thought of Brazil as an area analogous to Africa—that is, an area on the route to India where they would stop for trade or barter in indigenous products and slaves but not establish permanent settlements beyond an occasional trading post. The most commercially viable resource of Brazil in the first decades proved to be the item that gave the country its name, brazilwood, a tropical hardwood useful as a textile dye. As with Africa, the Portuguese government let out contracts for the trade to private individuals.

Brazilwood

The brazilwood industry did not bring about the founding of cities or other marks of full development, but its bulk was considerable for a time, and it was not a pure trade in natural products but involved some intervention on the part of the Portuguese. Though indigenous men of the region were accustomed to cutting down forest trees to clear fields, they did not have a tradition of commerce in trees, nor were they able to cut them on a large scale. The Portuguese therefore had to provide European axes and saws as well as product specifications. A Portuguese factor, or trading agent, would acquire the logs and have them ready when the ships came. Trading posts were often on islands, as in Africa, and a little later the first formal Portuguese settlements were also founded on islands. The only Portuguese who could be said to be actually settled in Brazil were some outcasts living among the Indians, who sometimes helped acquire useful Indian alliances.

About 1530 the Portuguese began to feel pressures to intensify their involvement with Brazil. Interlopers, especially the French, had begun to appear; the India trade was in a slump; and the great successes in Spanish America represented both an incentive and a threat. In response to such stimuli, the Portuguese sent an expedition to drive out the French and assert their authority. A number of settlers accompanied the expedition, which established the first formal Portuguese settlement—São Vicente—in 1532 on an island near present São Paulo.

The Portuguese had thus far acted entirely within their maritime-commercial tradition, and they continued for some time to do so, adopting measures quite different from those of the Spaniards. Whereas the Spaniards expanded from one area to the next in relay fashion, the Portuguese crown, in the mid-1530s, divided the entire Brazilian coast into strips of donatary captaincies, of which there were eventually 15. It granted them to *donatários,* prominent people presumed to have the personal resources to carry out the occupation and exploitation of their regions. The office was hereditary, with extensive judicial and administrative powers. The Portuguese had previously used this type of concession for their Atlantic island possessions. The encomienda, the master institution of 16th-century Spanish America, was not employed. From the first, though, leading Portuguese acquired large *sesmerias,* or land grants.

Donatary captaincies

In the event, several of the captaincies were never occupied at all, and others survived only for a short time. However, four of them led to permanent settlements, and two of these, São Vicente in the south and Pernambuco in the north, proved distinctly viable and profitable.

As on much of the Spanish fringe, the first Portuguese settlements in Brazil had to be fortified against Indian attacks. Provisioning was difficult, and for a time the Portuguese got much of their food through trade with the indigenous people, becoming accustomed to manioc (cassava) as their staple rather than wheat, which grew poorly in much of the region. Two types of agricultural establishments emerged: *roças,* which were food farms or truck gardens near towns, and fazendas, or export enterprises. The last were mainly sugar plantations, which were not yet very prosperous, even though conditions for sugar growing and transport were ideal in many places, because of lack of capital to build mills and buy African slave labour. The Portuguese at first tried to extract labour from the indigenous people in exchange for European products, but the effort failed, in part because the men of these semisedentary societies were not accustomed to agricultural labour. As had happened in Spanish America, the Brazilian settlers soon turned to Indian slavery for workers; slaves were acquired through raiding or through purchase from other Indians. A minority of more expensive African slaves formed a labour elite, much as in Spanish America.

Royal government

In 1548, still in response to much the same pressures and incentives as in 1530, the Portuguese decided to set up direct royal government in Brazil. The crown named a governor-general who took an expedition of a thousand people to Brazil, establishing a capital for the entire country in Bahia on the northeastern coast. In 1551 a bishopric was created. Thus it was not until 50 years after contact that Brazil achieved the level of institutionalization characteristic of the Spanish-American central areas almost from the beginning. The pace of development was much more comparable to that on the Spanish-American fringe.

At about this same time the Jesuits began to arrive, soon becoming the strongest arm of the church, as opposed to in Spanish America, where they arrived long after the other orders. They were prominent in the attempt to deal with the indigenous population, founding villages (*aldeias*) on new sites much in the manner of the missions on the Spanish-American fringe. Thus the main forms of European-Indian contact in Brazil—war, trade, slavery, and missions—were the same as on the periphery of Spanish America.

The Portuguese population in 16th-century Brazil remained sparse. Moreover, by all indications, including the Portuguese practice of exiling convicts to Brazil, one can imagine that it was as acutely marginal socially as were the settlers of Spanish-American fringe areas.

The sugar age. Starting in the last decades of the 16th century, the Brazilian sugar industry began an upswing that led to its being in the 17th century the world's largest producer of sugar for the ever-growing European market. The main structural changes had occurred by 1600, though the strongest growth came thereafter.

The more the industry prospered, the more it attracted Portuguese immigration, and the more it could afford African slaves as workers. Both movements resulted in the diminution of the indigenous role; by the third decade of the 17th century, through death and flight to the interior, Indians had become a negligible factor on the northeastern coast, where sugar growing concentrated. The Portuguese coming into the area were not only more numerous but represented a much broader cross section of society, including enough women for prominent men to marry. The northeastern cities were beginning to look more like their Spanish-American counterparts. In a word, the northeast was becoming a new central area, with some noticeable differences from those of Spanish America: it was built on bulk export rather than precious metals, with an Afro-European base rather than Indo-European, oriented to the sea rather than to an indigenous hinterland.

Sugar production was almost as industrial an enterprise as silver mining. The dominant feature was the *engenho,* the mill. So expensive were the mill, technicians' salaries, and the force of African slaves to work there that mill owners normally depended on cane growers called *lavradores* to produce cane for the mill. Under various kinds of leasing arrangements, the *lavradores* used their own African slave crews to cultivate the land, grow the cane, and transport it to the mill. Some of the cane growers were from mill-owning families, while others were more humble, and some even were racially mixed.

The sugar industry required a large number of Portuguese. Although Africans came to constitute the majority of the local population, the Portuguese sector was also large. Instead of a sprinkling of masters among great masses of slaves, the predominant pattern was the use of slaves in relatively small units, each in contact with some Portuguese. The mill owners had rural residences, but, as with the Spaniards, their main seats were in the nearest city, where their group tended to dominate the *senado da câmara,* the equivalent of the Spanish cabildo. Portuguese with less capital went into growing tobacco for export or *roças* for provisioning the cities and mills, and they employed relatively fewer slaves. In the backland (*sertão*), ranches grew up to supply the coast with meat and work animals. Society was varied and complex.

The rural-urban continuum was strong, and the Africans took part in it as well as the Portuguese, so that the most skilled and acculturated of them tended to end up in the cities, where there came to be an African population, increasingly racially mixed and in part free, much as in Spanish America. With so many more Africans present than in the Spanish central areas, groups based on African ethnicity could retain their language and cohesiveness longer. Christian lay organizations with an African ethnic base were very strong, and many African cultural elements were preserved, especially in the areas of music, dance, and popular religion. The same sort of strength allowed for the flourishing of independent communities of runaway slaves to an extent not known in Spanish America, though the phenomenon occurred there too in some forested areas.

Status system

An elaborate scaled status system recognizing racial and cultural mixture and legal status, comparable to the Spanish-American ethnic hierarchy, grew up in the Brazilian northeast, but it was different in being overwhelmingly bipolar—European and African—with the indigenous factor hardly counting. It is not by chance that in Mexico and Peru the top category remained Spaniard, while in Brazil it came to be white as well as Portuguese. If in the Spanish central areas the Africans were intermediaries, here they had a more complex function, replacing the Indians at the bottom of the functional ladder as well as filling many intermediate niches.

The northeast now assumed many of the other characteristics of a central area. The mercantile interest grew strong, localizing the form of men of business (*homens de negócios*) who both invested in merchandise and owned sugar mills. They intermarried with the planters and served on the town councils. Not only did a governor-general, later a viceroy, reside in Bahia, but there was (most of the time) a high court of appeal, or *relação,* like the Spanish-American audiencia, with the associated network of lawyers and notaries. Monasteries and convents became part of the picture, and authors writing on local topics appeared, some of the most prominent of them Jesuits.

Institutionalization stopped short of what was seen in the Spanish-American central areas, however. Transatlantic contact remained more essential to local society than in Spanish America. Universities and printing presses were not established; students went to Portugual for advanced education, and books were printed there. Transatlantic careers spanning not only Portugal and Brazil but also including Africa were common. So much a part of the Atlantic world was the northeast of Brazil that Europe continued to make itself felt strongly. It was perhaps a somewhat secondary phenomenon that the king of Spain was also the king of Portugal from 1580 to 1640, but the impact of the Netherlands was more directly felt, for the Dutch seized Bahia in 1624, holding it to 1625, and controlled the important captaincy of Pernambuco from 1630 to 1654.

The south. Only the northeast of Brazil was thoroughly transformed by the sugar industry. The remainder long stayed much as it had been before, a sparsely inhabited fringe with a weak economy, more indigenous and European in composition than African. São Paulo, the dominant centre of the south, had a small Portuguese population, and much if not most of it was racially mixed. Not unlike the Paraguayan Spaniards, the Paulistas (citizens of São Paulo) lived in large households and estates among numbers of Indian slaves, freedmen, and dependents, strongly affected by indigenous language, customs, diet, and family structure.

The products of the estates being in little demand elsewhere, much attention went into the area's most negotiable commodity, indigenous slaves. Desired at first to work on coastal plantations, Indian slaves lost marketability as the sugar industry was able to make the transition to Africans. But when the Dutch seized a part of the northeast and disrupted the African slave supply in the first half of the 17th century, the Paulistas' Indian slaves were more salable until the African supply lines were once again secured after mid-century. Thereafter the Paulistas turned more to exploring the interior, establishing new settlements there and searching for precious metals.

The *bandeira*

The Paulistas are known for an expeditionary form, the *bandeira* ("banner"), which, though by origin related to the conquering and exploring expeditions seen elsewhere, evolved almost beyond recognition and became a key element of Paulista culture. As time went on, it was necessary to go farther and farther for slaving, eventually to the areas of the Paraguayan Spaniards and even beyond. The *bandeirantes,* as the participants were called, might spend many months or even years in the backlands. Although led by Portuguese or people of mixed heritage passing for Portuguese, the highly mobile columns were mainly indigenous, being made up of direct dependents or slaves of the leaders or members of allied Indian groups. Though possessing some European weapons and cultural elements, they were highly adapted to the surroundings, using indigenous food, language, transportation, and much else. It was they above all who were responsible for making Brazil more than a coastal strip.

Spanish and Portuguese America in 1784.

Reprinted by permission of the publisher, from *People and Politics of Latin America,* copyright © 1955 by Ginn, 1945 by Ginn & Co., 1938 by Mary W. Williams, 1930 by Mary W. Williams; all rights reserved

SPANISH AMERICA IN THE AGE OF THE BOURBONS

A series of important changes occurring in Spanish America in the 18th century is often associated with dynastic changes in Spain—the replacement of the Habsburgs, who had ruled Spain since the early 16th century, by a branch of the French Bourbons in 1700. Little, however, altered in the Spanish territories until more than 50 years later, especially during the reign of Charles III (1759–88). Internal evolution and worldwide developments were doubtless more important in bringing about the new phenomena than the policy of a particular dynasty or ruler.

Economy and society. Demographic growth picked up sharply after about the mid-18th century in all areas about which information is available and in all sectors of the population. At the same time, economic activity increased in bulk, and prices rose steadily instead of fluctuating as they had been doing for centuries. Silver production, which was still at the base of the export economy of the old central areas, increased sharply, especially in Mexico, and so did the scale of operations and the input of capital, with strong participation by merchant-financiers. At the same time, local textile production had grown in size and economic importance, as demand rose in its market—humble Hispanized people in the city and countryside.

The large merchants had continued the process of localization to the point where only their birth was foreign; large firms tended to pass from a Spanish immigrant owner to his immigrant nephew. In every other way—marriage, investment, and residence pattern—the merchants were part of the local milieu, and, since export-import commerce was so important to the economy, they had risen to the top on the local scene; the wealthiest of them owned strings of haciendas in addition to their commercial and mining interests, and they acquired titles of high nobility.

Racial and cultural fusion

Racial and cultural fusion had advanced so far that the categorization embodied in the ethnic hierarchy could no longer capture it. Labels proliferated to designate complex mixtures, but the new terms sat lightly on those so labeled and often had no legal status. In everyday life, people who were able to function within a Hispanic context were often not labeled at all; many others changed almost at will from one category to another. One reaction to the excessive categorization was simplification, with only three categories—Spaniards, *castas,* and Indians—and often only two—Indians and others. The people of mixed descent were now so fully acculturated and so deeply embedded in local Hispanic society that they were qualified for and began to compete for nearly all positions except the very highest. There was, naturally, a reaction on the part of those most highly placed. With mulattoes entering the universities in numbers, ordinances began to declare that they were not eligible. With the children of wealthy Spaniards, humbler and racially mixed Spaniards, and *castas* all intermarrying widely, government and the church began to resist, declaring marriages between those differently labeled to be illegal and reinforcing the authority of parents in disallowing matches.

Such reactions did little to change the basic reality: the intermediate groups had grown and were continuing to grow to the extent that they could no longer be confined to their traditional intermediary functions. There were too many of them for all to become majordomos and artisans, and, in any case, many people called Indians by now could speak Spanish and handle tasks very well themselves for which intermediaries had previously been required. Since the people in the middle were no longer at a premium, their remuneration often decreased. If some pressed on into the higher strata, others were reduced to positions traditionally belonging to Indians, such as permanent labourer. In many areas the mixed groups were pouring into indigenous settlements at such a rate as to disrupt them and change their character.

Transformation of the east coast. Well into the 18th century, the perception that Mexico and Peru formed the centre and all the rest the periphery was still valid. By the last decades of the century, however, things were moving quickly in a different direction, favouring the Atlantic seaboard. European demand for tropical crops and even for temperate products, especially hides, increased substantially. At the same time, ships grew larger and faster. As a result, transatlantic shipment of bulk products became more viable, and trade routes shifted.

Argentina. The La Plata region had been very much on the edges of the Latin-American world since the conquest. The first founding of Buenos Aires in the early 16th century had failed, the survivors having taken refuge in the lands of the semisedentary Guaraní of Paraguay. The most developed area was the northwest interior, closest to the Potosí mining region, which supplied the mines with various products. Paraguay remained in relative isolation and poverty, participating in the money economy by sending its yerba maté (a tealike beverage) toward Peru. Buenos Aires was eventually refounded but remained a tiny, struggling port. The plains were inhabited by wild cattle (descendants of domestic animals introduced into the region earlier), nonsedentary Indians, and some highly localized mestizos later to be called gauchos.

Starting in the 1770s, improved transatlantic navigation transformed the region. Buenos Aires began to be able to compete with the older route through Panama and Peru in importing European goods for the mining region and exporting silver. The immigration of merchants and others increased. Taking advantage of the opportunity, the crown created the Viceroyalty of the Río de la Plata based in Buenos Aires (1776), including the Potosí mining region, which was taken from Peru. Buenos Aires became a capital with all the institutions associated with Lima or Mexico City. The city's population, including a substantial number of Africans because of its location on a slave route and its new wealth, grew explosively, and it began to exercise dominance over the interior, reversing the older scheme.

Buenos Aires

Yet Buenos Aires was not quite like Lima or Mexico City; it showed its newness, and traces of peripherality remained. The merchants of Buenos Aires had the same Spanish origins as their counterparts in Mexico City, but they were more closely tied to Spain, much like central-area merchants in the conquest period. They were more dominant locally, for there were no long-established families to compete with, and they came close to monopolizing the capital's municipal council. But they were far less wealthy than the largest Mexico City merchants, established no noble titles, and owned few or no rural estates. Indeed, there were no estates to buy: haciendas existed in the older northwestern region, but on the plains or pampas around Buenos Aires estate development had hardly begun. The hide export industry that now began to become prominent rested at first mainly on hunting wild animals; the merchants who exported hides were still secondary to those importing merchandise and exporting silver. Only in the last years before independence did merchants and others finally begin to build up estates and raise cattle in the more customary manner.

Venezuela. With its Caribbean coast, Venezuela had long been in a relatively favourable position in regard to potential availability of markets. By the 17th century the Caracas region was exporting cacao to Mexico, where most of the market for that product was then located, enabling it to begin buying African slaves for labour. As Europe joined the market and absorbed larger quantities of cacao by the late 18th century, Caracas became an urban centre comparable in size and institutionalization to Buenos Aires (though without a viceroy), and it had a better-developed hinterland of secondary settlements. The population along the coast was mainly European, African, and mixtures thereof. The situation, then, bore some similarity to that in Brazil.

The Caribbean islands. The Spaniards from the first had concentrated on the Greater Antilles, leaving the smaller islands virtually unoccupied. As developments passed the Spanish Caribbean by, even portions of the larger islands were left under-occupied. Thus, in the course of the 17th century, the French and English, aided by buccaneers of their respective nationalities, were able to take over the small islands, Jamaica, and the western end of Hispaniola to grow tropical crops, above all sugar, for themselves. The societies that grew up there were not exactly Latin-American in the usual sense; though in a way comparable to the society of northeastern Brazil, they were different in that the African slave population vastly outnumbered the Europeans, who were not only very few but also not well rooted, retaining intimate connections with the home countries. By the late 18th century the non-Spanish Caribbean islands had replaced Brazil as the world's greatest sugar producers.

The Spanish Caribbean islands (primarily Cuba and Puerto Rico) did not participate in the sugar boom, which was predicated on the notion of self-supply by the northern European nations. The population was more balanced between European and African than in the French and English possessions. In the second half of the 18th century the Cuban economy grew rapidly on the basis of tobacco export and provisioning of fleets and Spanish Caribbean ports. Only after the slave revolt in French Haiti in 1791, with great loss of French production, did Cuba begin to move in the direction of large-scale sugar export.

The Bourbon reforms. The Enlightenment, emanating to a large extent from France, penetrated both Spain (aided by the French origin of the Bourbons) and Spanish America in the 18th century. By the late part of the century individuals and organized societies in many of the American territories were producing journals and books in the manner of the work of the French Encyclopédistes, promoting reason, universality, science, modernity, and efficiency. Most Spanish-American writers, while staying in close touch with European currents, were concerned with the development, in practical terms, of their own regions.

Enlightenment philosophy bore importantly on government, which was called on to be more rationally unified, efficient, and free of church influence. Such ideas affected policy makers for the Spanish crown, and a series of activist royal measures of the 18th century were carried out in that spirit. Yet the timing and the nature of these moves had at least as much to do with changing conditions as with ideology. Most reforms came in a bundle in the late 18th century, the creation in 1739 of the Viceroyalty of New Granada based in Santa Fé (Bogotá) being an exception.

A major Bourbon reform, taking place mainly in the 1780s, was the creation of large districts called intendancies (the word and model were French). Each was headed by an official with extensive powers called an intendant, who was directly responsible to the crown in Spain. The measure was meaningful because royal government in the provinces, outside the seat of the governor or viceroy in each country, had hardly existed. It was as though a host of provincial cities received their own viceroy. One result, and indeed the one most intended, was an increase in revenue collection; another, not intended, was decentralization and bickering. The intendancy seats were not arbitrarily created or chosen but were mainly large cities that had once been encomendero centres and were still bishoprics, or long-lasting, large-scale mining centres. The change was realistic in that it recognized the immense growth and consolidation of provincial Hispanic centres that had occurred in the centuries since the first establishment of the viceroyalties, and for that reason it took hold. Less successful was the attempt to introduce similar officials at a lower level in the Indian countryside.

Intendancies

Military affairs were a second target of reform. Spanish America had long been defended by a patchwork of viceregal guards, port garrisons, half-fictional militias, and some forts and paid soldiers on frontiers with hostile Indians, but it had not had a formal military organization. In the late 18th century it acquired one, partly because of an increased foreign threat (Havana was occupied by the British in 1762–63), partly because the Bourbons imagined the army to be the most responsive branch available to them, and partly because professionalization of the military was an international trend of the time. A relatively small number of regular units formed the backbone for a larger, more rigorously organized militia. At first the regulars were brought in from Spain, but before long the lower ranks were mainly locals, and locals found entry even into the officer ranks, though the top commanders were usually Spaniards born. The military was primarily Hispanic, with Indians taking part only under exceptional circumstances, and it reflected local society, with officers drawn from prominent families and many persons of mixed descent and Africans among the enlisted men. Organized in local districts, the units' loyalties were above all local as well.

Government in Bourbon times was not antireligious, but it was sufficiently affected by the spirit of the times to be quite anticlerical. The most decisive of the measures taken was the expulsion of the Jesuit order from Spanish America and Spain in 1767. Preceded by similar actions in Portugal and France, the move was part of an international wave, but it also made excellent sense in purely Spanish-American terms. Although the Jesuits were the wealthiest of the orders, they had arrived last, had fierce rivals in other branches of the church, and counted few locals among their members. Thus their expulsion was greeted with (usually hidden) approval by many. The crown in general tried to further the secular clergy over the religious orders (imagined to be more independent-minded), but the policy had little effect except in areas where the secular clergy, which grew with the expansion of civil society, was already on the rise. Almost on the eve of independence, the crown attempted to confiscate church property, but the measure proved hard to enforce.

The late Bourbons favoured more active encouragement of the economy and even intervention in it. They provided tax reductions and technical aid for the silver mining industry; they expanded state monopolies beyond the mercury needed for mining to some other commodities, of which tobacco was the most successful. Their largest reform, however, went in the opposite direction, consisting in the declaration of free trade within the Spanish empire, so that any port could trade with any other at will.

Imperial free trade

In earlier times the bulk of transatlantic trade had been directed at Mexico and Peru, and annual convoys sponsored by the Spanish government were an efficient way not only to organize the traffic but also to protect it from pirates, who were the main threat. By the 18th century the northern European powers had naval superiority and could easily have destroyed any convoy. Moreover, in Spanish America new central areas had arisen, with a consequent diversification of destinations, and in Spain the north had revived at the expense of the south, where Seville and Cádiz had monopolized Indies navigation. Under these changed circumstances, the best arrangement was to allow individual ships to travel between any Spanish port and any American port. The fleet system gradually fell apart in the 18th century. Imperial free trade was introduced between 1765 and 1789, first affecting Cuba and spreading to all Spanish possessions. The measure coincided with a marked increase in commercial volume; that this itself caused the increase, as opposed to demographic growth in the Indies and industrial growth in Europe, is not clear. Nor are the effects entirely clear. The deluge of goods made it harder for the largest American merchants to be as dominant as previously, and for the first time local textile producers had real competition for the lower end of the market. Even so, the large firms of Mexico City were not destroyed, and the Puebla textile industry continued to grow.

BRAZIL AFTER 1700

In the late 17th century the explorations of the Paulistas finally led to the discovery of major gold deposits in a large district inland from Rio de Janeiro that became known as Minas Gerais. As the news spread, outsiders poured into the area. A time of turbulence, with the frontier Paulistas trying to assert their rights, ended after a few decades with the victory of the newcomers and the entry of royal authority. The south-centre, both the coast and the near interior, now took on the essential characteristics of the northeast—of a land living on European exports and inhabited by a population mainly Portuguese, African, and mulatto, with a large sector of slaves, along with many recently freed persons. The mining district flourished during the time of the boom, generating a network of settlements where none had been before and a local culture that included the now-renowned architectural style of its small churches.

More importantly for Brazil as a whole, Rio de Janeiro began to become an important urban centre in the usual mold, and the institutional component thickened, just as it had earlier on the basis of mineral wealth in the old Spanish-American central areas. By 1763 Rio had become the capital of Brazil, replacing Salvador in the northeast. Although the northeastern sugar industry continued to export more by value than the gold region, the latter had newer wealth and perhaps a higher profitability, and distant regions began to orient themselves to it in important ways. Stock-raising regions both in the northern interior and on the southern plains sent their animals to the mines, thereby both growing and helping unify the country.

The chronology of Brazil does not mesh closely with that of Spanish America in the late period. The gold boom was a type of development that had occurred much earlier in the Spanish territories; moreover, it did not last into the second half of the 18th century, when the most marked economic growth was occurring elsewhere, but began to decline by mid-century. Brazil had already experienced the bulk export revolution in the 17th century with sugar, and in the later 18th century exports were actually declining much of the time. Some growth, however, occurred late in the century in response to the decline of the French sugar industry in the Caribbean after the slave revolt in Haiti and some experimentation with new crops that were beginning to be of interest in Europe. Thus, though the Portuguese were as much affected by the Enlightenment as the Spaniards and had their time of active reform under the marquês de Pombal, prime minister and in effect ruler of Portugal in the period 1750–77, the context was hardly comparable. Outstanding among the actions taken under his ministry was a wave of expulsions of the Jesuits, in 1759. During his long rule Pombal instituted numerous fiscal and administrative reforms and even attempted social legislation. He gave much attention to the far north of Brazil, attempting to develop the region, and a time of considerable local development and change did in fact coincide with his activity.

Marquês de Pombal

PREINDEPENDENCE PHENOMENA

The position of the locally born Spaniards, often called creoles or *criollos* (though they were slow to call themselves that), had been growing stronger all across the post-conquest centuries. From an early time they owned most of the rural estates and dominated most of the cabildos. By the 17th century they were a large majority among the secular clergy and prominent in the orders, and as time went on they received more and more of the bishoprics. In the course of the 17th century they achieved appointments as audiencia judges in various centres, and by the second half of the 18th century they were dominating, sometimes virtually monopolizing, the membership of audiencias all over Spanish America. As the military came into existence, they found prominent places in it. Large mining producers might be either born locally or Spanish-born. Large merchants remained predominantly born in Spain, but they married into local families, whose interests they often served. Each major local Spanish family had members placed strategically across the whole system, creating a strong informal network. Only the viceroys and usually the archbishops were normally recruited from the outside, and even they had local entourages.

As the Bourbon government in Spain became more active late in the 18th century, it wanted a larger place for its own Spanish-born associates and began to view the extent of local American dominance with alarm. The audiencias were gradually filled predominantly with Spanish-born judges; nearly all the intendants were outsiders and so were the highest military officers. Yet the basic situation hardly changed, for the Spanish-born appointees had to function in a local milieu, into which they were rapidly absorbed. As independence approached, the local Spaniards or creoles had influence and experience at all levels of society, economy, and government, but they had been under challenge for a generation or more and were correspondingly resentful.

Consciousness of separateness of various kinds had been growing for a long time. In Mexico, starting as early as the mid-17th century, the illustrious indigenous past and the cult of the Virgin of Guadalupe had become a basis for national pride, promoted above all by creole priests and scholars. Other areas had approximate equivalents, if not as well-defined. Awareness of ethnic distinctions within the Spanish category increased in the 18th century along with the proliferation of ethnic terminology in general. The creoles were still mainly called Spaniards, but the new arrivals from Spain, now a small minority, were distinguished from the rest as peninsular or European Spaniards, and in Mexico they received the insulting nickname *gachupín*.

Awareness of separateness

The middle groups, whether humble Spaniards or people in the racially mixed categories, had much reason for discontent. The expansion of the middle left a large segment of the population without employment corresponding to its expectations and capacities. Corporately organized indigenous groups, however, though not in an admirable state economically or in many other respects, were generally little concerned about conditions at a countrywide level. It is not that they were apathetic; all through the intervening centuries they had stood up for themselves, through litigation and sometimes through disturbances and revolts, but they had done so as individual communities. On the nonsedentary fringe, wars and rebellions continued, but this was not different from earlier times. The most volatile element were Spanish-speaking Indians in and around Hispanic communities, who had mobility and broad awareness and whose profile no longer corresponded to the implications and duties of the label "Indian."

Two large manifestations of the late 18th century can be seen as foreshadowing independence, though it is possible that they did as much to retard it. In 1780–81 the Andean highlands experienced the Tupac Amaru revolt, which wrested control of much of the region from the ordinary authorities for many months until it was forcibly put down. Although references were made to the Inca heritage and the rebellion was indeed based in the indigenous countryside, its leaders were largely provincial mestizos and creoles from the middle levels of local society. The so-called Comunero Rebellion in Colombia began in 1780 in the provincial town of Socorro, a tobacco and textile producing centre. From there it spread widely before disbanding a year later largely as a result of negotiations.

Both movements were in immediate response to Bourbon fiscal measures, and both proclaimed ultimate loyalty to the Spanish crown. In Peru especially, there was a strong reaction afterward against both dissent and the indigenous population. The impetus for independence in Spanish South America would eventually come from the newly thriving Atlantic seaboard regions—the former fringes, Venezuela and Argentina—which had mobile Hispanized populations and lacked large groups of sedentary Indians. In Mexico too, things would start in the very similar near north of the country.

The situation in Brazil

In Brazil, the local Portuguese population had a position quite comparable to that of the Spanish-American creoles, but it was not so far advanced, and the situation had not become polarized. Transatlantic mobility still made itself felt, with many leading Brazilian Portuguese having been educated in Portugal. Locally born Portuguese had long participated in the Brazilian high court system, but they had never been a majority as in Spanish America. Two well-known rebellious incidents occurring in the 1780s and '90s, in Minas Gerais and Bahia, did not have full support even locally.

Latin America approached independence after a thoroughgoing ethnic and cultural transformation across a period of over three centuries. That process did not destroy the indigenous component, which was still very much alive corporately and culturally in the old central areas and some other regions and had also affected and entered into the mixed Iberian societies that had come to dominance. Even where it almost disappeared, the indigenous factor was important, for its weakness or absence was what allowed certain regions to become more European and African. Most of the independent countries that arose in the early 19th century went back to indigenous culture areas that had been re-formed into functional units under Iberian management in the 16th century. (J.Lo.)

The independence of Latin America

After three centuries of colonial rule, independence came rather suddenly to most of Spanish and Portuguese America. Between 1808 and 1826 all of Latin America except the Spanish colonies of Cuba and Puerto Rico slipped out of the hands of the Iberian powers who had ruled the region since the conquest. The rapidity and timing of that dramatic change were the result of a combination of long-building tensions in colonial rule and a series of external events.

The reforms imposed by the Spanish Bourbons in the 18th century provoked great instability in the relations between the rulers and their colonial subjects in the Americas. Many Creoles (those of Spanish parentage but who were born in America) felt Bourbon policy to be an unfair attack on their wealth, political power, and social status. Others did not suffer during the second half of the 18th century; indeed, the gradual loosening of trade restrictions actually benefited some Creoles in Venezuela and certain areas that had moved from the periphery to the centre during the late colonial era. However, those profits merely whetted those Creoles' appetites for greater free trade than the Bourbons were willing to grant. More generally, Creoles reacted angrily against the crown's preference for peninsulars in administrative positions and its declining support of the caste system. After hundreds of years of proven service to Spain, the American-born elites felt that the Bourbons were now treating them like a recently conquered nation.

Enlightenment ideas

In cities throughout the region, Creole frustrations increasingly found expression in ideas derived from the Enlightenment. Imperial prohibitions proved unable to stop the flow of potentially subversive English, French, and North American works into the colonies of Latin America. Creole participants in conspiracies against Portugal and Spain at the end of the 18th and beginning of the 19th centuries showed familiarity with such European Enlightenment thinkers as Thomas Hobbes, John Locke, Montesquieu, and Jean-Jacques Rousseau. The Enlightenment clearly informed the aims of dissident Creoles and inspired some of the later, great leaders of the independence movements across Latin America.

Still, these ideas were not, strictly speaking, causes of independence. Creoles selectively adapted rather than simply embraced the thought that had informed revolutions in North America and France. Leaders in Latin America tended to shy away from the more socially radical European doctrines. Moreover, the influence of those ideologies was sharply restricted; with few exceptions only small circles of educated, urban elites had access to Enlightenment thought. At most, foreign ideas helped foster a more questioning attitude toward traditional institutions and authority.

European diplomatic and military events provided the final catalyst that turned Creole discontent into full-fledged movements for Latin-American independence. When the Spanish crown entered into an alliance with France in 1795, it set off a series of developments that opened up economic and political distance between the Iberian countries and their American colonies. By siding with France, Spain pitted itself against England, the dominant sea power of the period, which used its naval forces to reduce and eventually cut communications between Spain and the Americas. Unable to preserve any sort of monopoly on trade, the Spanish crown was forced to loosen the restrictions on its colonies' commerce. Spanish Americans now found themselves able to trade legally with other colonies, as well as with any neutral countries such as the United States. Spain's wartime liberalization of colonial trade sharpened Creoles' desires for greater economic self-determination.

Occurrences in Europe in the early 19th century created a deep political divide between Spain and its American colonies. In 1807 the Spanish king, Charles IV, granted passage through Spanish territory to Napoleon, who was on his way to invade Portugal. The immediate effect of that concession was to send the Portuguese ruler, Prince Regent John, fleeing in British ships to Brazil. Arriving in Rio de Janeiro with some 15,000 officials, nobles, and other members of his court, John transformed the Brazilian colony into the administrative centre of his empire. When Napoleon turned on his Spanish allies in 1808, events took a disastrous turn for Spain and its dominion in the Americas. Shortly after Charles had abdicated in favour of his son Ferdinand, Napoleon had them both imprisoned. With these figures of legitimate authority in his power, the French ruler tried to shatter Spanish independence. In the process he set off a political crisis that swept across both Spain and its possessions. The Spanish political tradition centred on the figure of the monarch, yet, with Charles and Ferdinand removed from the scene, the hub of all political authority was missing.

In 1810 a Cortes (Parliament) emerged in Cádiz to represent both Spain and Spanish America. Two years later it produced a new, liberal constitution that proclaimed Spain's American possessions to be full members of the kingdom and not mere colonies. Yet the Creoles who participated in the new Cortes were denied equal representation. Moreover, the Cortes would not concede permanent free trade to the Americans and obstinately insisted on reasserting its control over Creoles. Having had a taste of freedom during their political and economic isolation from the mother country, Spanish Americans did not easily consent to a reduction of their power and autonomy.

Two other European developments further dashed the hopes of Creoles, pushing them more decisively toward independence. The year 1814 saw the restoration of Ferdinand to the throne and with it the energetic attempt to reestablish Spanish imperial power in the Americas. Rejecting compromise and reform, Ferdinand resorted to military force to bring wayward Spanish-American regions back into the empire as colonies. The effort only served to harden the position of Creole rebels. In 1820 troops waiting in Cádiz to be sent as part of the crown's military campaigns revolted, forcing Ferdinand to agree to a series of liberal measures. That concession divided and weakened loyalist opposition to independence in the Americas. Many supporters of the crown now had doubts about the monarchy for which they were fighting.

The wars of independence, 1808–26

The final victory of Latin-American patriots over Spain and the fading loyalist factions began in 1808 with the political crisis in Spain. With the Spanish king and his son Ferdinand taken hostage by Napoleon, Creoles and peninsulars began to jockey for power across Spanish America. During 1808–10 juntas emerged to rule in the name of Ferdinand VII. In Mexico City and Montevideo caretaker governments were the work of loyal peninsular Spaniards eager to head off Creole threats. In Santiago, Caracas, Bogotá, and other cities, by contrast, it was Creoles who controlled the provisional juntas. Not all of these governments lasted very long; loyalist troops quickly put down Creole-dominated juntas in La Paz and Quito. By 1810, however, the trend was clear. Without denouncing Ferdinand, Creoles throughout most of the region were moving toward the establishment of their own autonomous governments. Transforming these early initiatives into a break with Spanish control required tremendous sacrifice. Over the next decade and a half, Spanish Americans had to defend with arms their movement toward independence.

SPANISH AMERICA

The southern movement in South America. The movements that liberated Spanish South America arose from opposite ends of the continent. From the north came the movement led most famously by Simón Bolívar, a dynamic figure known as the Liberator. From the south proceeded another powerful force, this one directed by the more circumspect José de San Martín. After difficult conquests of their home regions, the two movements spread the cause of independence through other territories, finally meeting on the central Pacific coast. From there, troops under northern generals finally stamped out the last vestiges of loyalist resistance in Peru and Bolivia by 1826.

The struggles that produced independence in the south began even before Napoleon's invasion of Portugal and Spain. In 1806 a British expeditionary force captured Buenos Aires. When the Spanish colonial officials proved ineffective against the invasion, a volunteer militia of Creoles and peninsulars organized resistance and pushed the British out. In May 1810 prominent Creoles in Buenos Aires, having vied with peninsulars for power in the intervening years, forced the last Spanish viceroy there to consent to a *cabildo abierto*, an extraordinary open meeting of the municipal council and local notables. Although shielding itself with a pretense of loyalty to Ferdinand, the junta produced by that session marked the end of Spanish rule in Buenos Aires and its hinterland. After its revolution of May 1810, the region was the only one to resist reconquest by loyalist troops throughout the period of the independence wars.

Buenos Aires

Independence in the former Viceroyalty of the Río de la Plata, however, encountered grave difficulties in the years after 1810. Central authority proved unstable in the capital city of Buenos Aires. An early radical liberal government dominated by Mariano Moreno gave way to a series of triumvirates and supreme directors. More troubling still were the bitter rivalries emerging between Buenos Aires and other provinces. From the start Buenos Aires' intention of bringing all the former viceregal territories under its control set off waves of discord in the outlying provinces. At stake was not only political autonomy per se but also economic interest; the Creole merchants of Buenos Aires, who initially sought the liberalization of colonial restraints on commerce in the region, subsequently tried to maintain their economic dominance over the interior. A constituent assembly meeting in 1813 adopted a flag, anthem, and other symbols of national identity, but the apparent unity disintegrated soon afterward. This was evident in the assembly that finally proclaimed independence in 1816; that body received no delegates from several provinces, even though it was held outside Buenos Aires, in the interior city of Tucumán (in full, San Miguel de Tucumán).

Distinct interests and long-standing resentment of the viceregal capital led different regions in the south to pursue separate destinies. Across the Río de la Plata from Buenos Aires, Montevideo and its surroundings became the separate Estado Oriental ("Eastern State," later Uruguay). Caught between the loyalism of Spanish officers and the imperialist intentions of Buenos Aires and Portuguese Brazil, the regional leader José Gervasio Artigas formed an army of thousands of gauchos. By 1815 Artigas and this force dominated Uruguay and had allied with other provinces to oppose Buenos Aires.

Buenos Aires achieved similarly mixed results in other neighbouring regions, losing control of many while spreading independence from Spain. Paraguay resisted Buenos Aires' military and set out on a path of relative isolation from the outside world. Other expeditions took the cause to Upper Peru, the province that would become Bolivia. After initial victories there, the forces from Buenos Aires retreated, leaving the battle in the hands of local Creole, mestizo, and Indian guerrillas. By the time Bolívar's armies finally completed the liberation of Upper Peru (then renamed in the Liberator's honour), the region had long since separated itself from Buenos Aires.

The southern independence forces

The main thrust of the southern independence forces met much greater success on the Pacific coast. In 1817 San Martín, a Latin-American-born former officer in the Spanish military, directed 5,000 men in a dramatic crossing of the Andes and struck at a point in Chile where loyalist forces had not expected an invasion. In alliance with Chilean patriots under the command of Bernardo O'Higgins, San Martín's army restored independence to a region whose highly factionalized junta had been defeated by royalists in 1814. With Chile as his base, San Martín then faced the task of freeing the Spanish stronghold of Peru. After establishing naval dominance in the region, the southern movement made its way northward. Its task, however, was formidable. Having benefited from colonial monopolies and fearful of the kind of social violence that the late-18th-century revolt had threatened, many Peruvian Creoles were not anxious to break with Spain. Consequently, the forces under San Martín managed only a shaky hold on Lima and the coast. Final destruction of loyalist resistance in the highlands required the entrance of northern armies.

The north and the culmination of independence. Independence movements in the northern regions of Spanish South America had an inauspicious beginning in 1806. The small group of foreign volunteers that the Venezuelan revolutionary Francisco de Miranda brought to his homeland failed to incite the populace to rise against Spanish rule. Creoles in the region wanted an expansion of the free trade that was benefiting their plantation economy. At the same time, however, they feared that the removal of Spanish control might bring about a revolution that would destroy their own power.

Creole elites in Venezuela had good reason to fear such a possibility, for a massive revolution had recently exploded in the French Caribbean colony of Saint-Domingue. Beginning in 1791, a massive slave revolt sparked a general insurrection against the plantation system and French colonial power. The rebellion developed into both a civil war, pitting blacks and mulattos against whites, and an international conflict, as England and Spain supported the white plantation owners and rebels, respectively. By the first years of the 19th century, the rebels had shattered what had been a model colony and forged the independent nation of Haiti. Partly inspired by those Caribbean events, slaves in Venezuela carried out their own uprisings in the 1790s. Just as it served as a beacon of hope for the enslaved, Haiti was a warning of everything that might go wrong for elites in the cacao-growing areas of Venezuela and throughout slave societies in the Americas.

Creole anxieties also contributed to the persistence of a strong loyalist faction in the Viceroyalty of New Granada, but they did not prevent the rise of an independence struggle there. Creoles organized revolutionary governments that proclaimed social and economic reforms in 1810 and openly declared a break with Spain the following year. Forces loyal to Spain fought the patriots from the start, leading to a pattern much like that which characterized the Plata: patriot rebels held the capital city and its surroundings but could not dominate large sections of the countryside. Some interpreted an 1812 earthquake that wreaked particular destruction on patriot-held areas as a sign of divine displeasure with the rebels. The year 1812 certainly was the onset of a difficult period for the independence armies of New Granada. Loyalist forces crushed the rebels' military, driving Bolívar into temporary exile.

The Liberator soon returned with a new army, and the battle entered a tremendously violent phase. After much of the local aristocracy had abandoned the independence cause, blacks and mulattos carried on the struggle. Elites reacted to the efforts of these common people with open distrust and opposition. Bolívar's forces invaded Venezuela in 1813, waging a campaign with a ferocity that is captured perfectly by their motto, "guerra a muerte" ("war to the death"). With loyalists displaying the same passion and violence, the rebels achieved only short-lived victories. The army led by the loyalist José Tomás Boves here demonstrated the key military role that the llaneros (cowboys) came to play in the region's struggle. Turning the tide against independence, these highly mobile, ferocious fighters made up a formidable military force that pushed Bolívar out of his home country once more.

"War to the death"

By 1815 the independence movements in Venezuela and almost all across Spanish South America seemed moribund. A large military expedition sent by Ferdinand VII in that year reconquered Venezuela and most of New Granada. Yet another invasion led by Bolívar in 1816 failed miserably.

The following year a larger and revitalized independence movement emerged, winning the struggle in the north and taking it into the Andean highlands. The mercurial Bolívar, the scion of an old aristocratic Creole family in Caracas, galvanized this initiative. Hero and symbol of South American independence, Bolívar did not produce victory by himself, of course; still, he was of fundamental importance to the movement as an ideologue, military leader, and political catalyst. In his most famous writing, the "Jamaica Letter" (composed during one of his periods of exile, in 1815), Bolívar affirmed his undying faith in the cause of independence, even in the face of the patriots' repeated defeats. While laying out sharp criticisms of Spanish colonialism, the document also looked toward the future. For Bolívar, the only path for the former colonies was the establishment of autonomous, centralized republican government. Liberal in some respects, the Liberator in the Jamaica Letter and elsewhere showed himself to be at heart socially conservative and politically authoritarian. Although he favoured the granting of many civil liberties to all male citizens and the abolition of slavery, for instance, Bolívar worried that the death of so many white soldiers in the wars would condemn Latin America to a system of "pardocracy," or rule by pardos (people of mixed-race ancestry, generally those of a black and white background), an outcome he obviously dreaded. The type of republic that he eventually espoused was very much an oligarchic one, with property qualifications on suffrage and power centred in the hands of a strong executive.

The Liberator emerged as a strong military and political force in the struggles that began in 1817. Bolívar at this point expanded the focus of the movement, shifting his attention to New Granada and courting supporters among the casta majority. A mixed-race group of llaneros led by José Antonio Páez proved crucial to the patriots' military victories in 1818–19. A major step in that success came in the subduing of the loyalist defenders of Bogotá in 1819. After leading his army up the face of the eastern Andes, Bolívar dealt a crushing defeat to his enemies in the Battle of Boyacá.

Consolidating victory in the north proved difficult. A congress that Bolívar had convened in Angostura in 1819 named the Liberator president of Gran Colombia, a union of what are today Venezuela, Colombia, Panama, and Ecuador. In reality, sharp divisions permeated the region even before Angostura; these ultimately dashed Bolívar's hopes of uniting the former Spanish colonies into a single new nation. The Bogotá area, for example, had previously refused to join in a confederation with the rest of revolutionary New Granada. Furthermore, loyalist supporters still held much of Venezuela, parts of the Colombian Andes, and all of Ecuador. Still, the tide had turned in favour of independence, and further energetic military campaigns liberated New Granada and Venezuela by 1821. A constituent congress held that year in Cúcuta chose Bolívar president of a now much more centralized Gran Colombia.

Leaving his trusted right-hand man, Francisco de Paula Santander, in Bogotá to rule the new government, Bolívar then pushed on into Ecuador and the central Andes. There the southern and northern armies came together in a pincer movement to quash the remaining loyalist strength. In 1822 San Martín and Bolívar came face-to-face in a celebrated but somewhat mysterious encounter in Guayaquil, Ecuador. Accounts of their meeting vary widely, but apparently San Martín made the realistic evaluation that only Bolívar and his supporters could complete the liberation of the Andes. From that point on, the northerners took charge of the struggle in Peru and Bolivia. After standing by while Spanish forces threatened to recapture all the lands that San Martín's armies had emancipated, Bolívar responded to the calls of Peruvian Creoles and guided his soldiers to victory in Lima. While he organized the government there, his lieutenants set out to win the highlands of Peru and Upper Peru. One of them, the Venezuelan Antonio José de Sucre, directed the patriots' triumph at Ayacucho in 1824, which turned out to be the last major battle of the war. Within two years independence fighters mopped up the last of loyalist resistance, and South America was free of Spanish control.

Mexico and Central America. The independence of Mexico, like that of Peru, the other major central area of Spain's American empire, came late. As was the case in Lima, Mexican cities had a powerful segment of Creoles and peninsular Spaniards whom the old imperial system had served well. Mexican Creoles, like those in Peru, had the spectre of a major social uprising to persuade them to cling to Spain and stability for a while longer. For many of the powerful in Mexican society, a break with Spain promised mainly a loss of traditional status and power and possibly social revolution.

What was unique to the Mexican case was that the popular rebellion that exploded in 1810 was actually the first major call for independence in the region. Between 1808 and 1810, peninsulars had acted aggressively to preserve Spain's power in the region. Rejecting the notion of a congress that would address the question of governance in the absence of the Spanish king, leading peninsulars in Mexico City deposed the viceroy and persecuted Creoles. They then welcomed weaker viceroys whom they knew they could dominate. Peninsulars' efforts could not, however, prevent the emergence of an independence struggle. In 1810 the Bajío region produced a unique movement led by a radical priest, Miguel Hidalgo y Costilla. When officials discovered the conspiracy that Hidalgo and other Creoles had been planning in Querétaro, the priest appealed directly to the indigenous and mestizo populace. A rich agricultural and mining zone, the Bajío had recently undergone difficult economic times that hit those rural and urban workers particularly hard. Thus many of them responded eagerly to Hidalgo's famous Grito de Dolores ("Cry of Dolores"). Although framed as an appeal for resistance to the peninsulars, the Grito was in effect a call for independence.

Grito de Dolores

The enthusiasm that Hidalgo stirred among Indians and mestizos shocked and frightened both Creole and peninsular elites. Under the banner of the Virgin of Guadalupe, the movement's ranks swelled rapidly. Hidalgo's untrained army grew to have some 80,000 members as it conquered towns and larger cities and ultimately threatened Mexico City itself. During their campaign the members of this force attacked the persons and property of peninsular and Creole elites. The movement for independence was becoming a race and class war.

Perhaps fearing the atrocities his troops might commit there, Hidalgo prevented the movement from entering Mexico City. Shortly afterward troops of the viceregal government caught up with the rebels. After a dramatic military defeat, Hidalgo was captured in early 1811 and executed.

The death of its first leader did not mean the end of Mexico's first independence campaign. Soon another priest, the mestizo José María Morelos y Pavón, took over the reins of the movement. Under Morelos the rebellion gained clearer objectives of independence and social and economic reform as well as greater organization and a wider social base. With the defeat and death of Morelos in 1815, the potential national scope of the movement came to an effective end. Although smaller forces under leaders like Vicente Guerrero and Guadalupe Victoria (Manuel Félix Fernández) continued to harass the powerful through guerrilla warfare in several regions, the popular movement for independence in Mexico was no longer a grave threat to elite power.

Final independence, in fact, was not the result of the efforts of Hidalgo, Morelos, or the forces that had made up their independence drive. It came instead as a conservative initiative led by military officers, merchants, and the Roman Catholic church. The liberals who carried out the 1820 revolt in Spain intended to eliminate the special privileges of the church and the military. Anxious over that threat to the strength of two of the pillars of the Mexican government and newly confident in their ability to keep popular forces in check, Creoles turned against Spanish rule in 1820–21.

Two figures from the early rebellion played central roles in liberating Mexico. One, Guerrero, had been an insurgent chief; the other, Agustín de Iturbide, had been an officer in the campaign against the popular independence movement. The two came together behind an agreement known as the Iguala Plan. Centred on provisions of independence, respect for the church, and equality between Mexicans and peninsulars, the plan gained the support of many Creoles, Spaniards, and even former rebels. As royal troops defected to Iturbide's cause, the new Spanish administrator was soon forced to accept the inevitability of Mexican independence. A year later, in 1822, Iturbide engineered his own coronation as Agustín I, Emperor of Mexico.

Iguala Plan

The following year, a revolt that included the former insurgent Guadalupe Victoria (who, like Guerrero, had abandoned the cause of a popular independence) cut short Iturbide's tenure as monarch. The consequences of that overthrow extended from Mexico through Central America. In Mexico the rebellion ushered in a republic and introduced Antonio López de Santa Anna, who occupied a central place in the nation's politics for several decades. The provinces of the Kingdom of Guatemala—which included what are today the Mexican state of Chiapas and the nations of Guatemala, El Salvador, Honduras, Nicaragua, and Costa Rica—had adhered to Iturbide's Mexico by 1822. With the exception of Chiapas, these Central American provinces split off from Mexico in the wake of Iturbide's fall. They formed a federation, the United Provinces of Central America, which held together only until 1838, when regionalism led to the creation of separate countries in the region.

BRAZIL

Brazil gained its independence with little of the violence that marked similar transitions in Spanish America. A conspiracy against Portuguese rule in 1789–98 showed that some groups in Brazil had already been contemplating the idea of independence in the late 18th century. Moreover, the Pombaline reforms of the second half of the 18th century, Portugal's attempt to overhaul the administration of its overseas possessions, were an inconvenience to many in the colony. Still, the impulse toward independence was less powerful in Brazil than in Spanish America. Portugal, with more limited financial, human, and military resources than Spain, had never ruled its American subjects with as heavy a hand as its Iberian neighbour. Portugal neither enforced commercial monopolies as strictly nor excluded the American-born from administrative positions as widely as did Spain. Many Brazilian-born and Portuguese elites had received the same education, especially at the University of Coimbra in Portugal. Their economic interests also tended to overlap. The reliance of the Brazilian upper classes on African slavery, finally, favoured their continued ties to Portugal. Plantation owners depended on the African slave trade, which Portugal controlled, to provide workers for the colony's main economic activities. The size of the resulting slave population—approximately half the total Brazilian population in 1800—also meant that Creoles shied away from political initiatives that might mean a loss of their own control of their social inferiors.

The key step in the relatively bloodless end of colonial rule in Brazil was the transfer of the Portuguese court from Lisbon to Rio de Janeiro in 1808. The arrival of the court transformed Brazil in ways that made its return to colony status impossible. The unprecedented concentration of economic and administrative power in Rio de Janeiro brought a new integration to Brazil. The emergence of that capital as a large and increasingly sophisticated urban centre also expanded markets for Brazilian manufactures and other goods. Much more important in this regard were two of Prince Regent John's first acts in Brazil. In 1808 he decreed the removal of old restrictions on manufacturing in Brazil and the opening of Brazilian ports to direct trade with all friendly nations. From that point on Brazil became more a metropolis than a colony.

Brazil headed into a political crisis when groups in Portugal tried to reverse the metropolitanization of their former colony. With the end of the Napoleonic Wars came calls for John to return to Lisbon. At first he demurred and in 1815 even raised Brazil to the status of kingdom, legally equal to Portugal within the empire that he ruled. The situation was a difficult one for John (after 1816 King John VI). If he moved back to Lisbon, he might lose Brazil, but if he remained in Rio, he might well lose Portugal. Finally, after liberal revolts in Lisbon and Oporto in 1820, the Portuguese demands became too strong for him to resist. In a move that ultimately facilitated Brazil's break with Portugal, John sailed for Lisbon in 1821 but left his son Dom Pedro behind as prince regent. It was Dom Pedro who, at the urging of local elites, oversaw the final emergence of an independent Brazil.

Dom Pedro

Matters were pushed toward that end by Portuguese reaction against the rising power of their former colony. Although the government constituted by the liberals after 1820 allowed Brazilian representation in a Cortes, it was clear that Portugal now wanted to reduce Brazil to its previous colonial condition, endangering all the concessions and powers the Brazilian elite had won. By late 1821 the situation was becoming unbearable. The Cortes now demanded that Dom Pedro return to Portugal. As his father had advised him to do, the prince instead declared his intention to stay in Brazil in a speech known as the "Fico" ("I am staying"). When Pedro proclaimed its independence on Sept. 7, 1822, and subsequently became its first emperor, Brazil's progression from Portuguese colony to autonomous country was complete.

Independence did not come without a price. Over the next 25 years Brazil suffered a series of regional revolts, some lasting as long as a decade and costing tens of thousands of lives. Dom Pedro I was forced from his throne in 1831, to be succeeded by his son, Dom Pedro II. The break with Portugal did not itself, however, produce the kind of disruption and devastation that plagued much of the former Spanish America. With its territory and economy largely intact, its government headed by a prince of the traditional royal family, and its society little changed, Brazil enjoyed continuities that made it extraordinarily stable in comparison with most of the other new states in the region.

Building new nations, 1826–50

While Brazil maintained its territorial integrity after independence, the former Spanish America split into more than a dozen separate countries, following the administrative divisions of the colonial system. The difficulty for the inhabitants of these units was not, however, as simple as the demarcation of geographic boundaries. Rather, the recently emancipated countries of Latin America faced the much more daunting challenge of defining and consolidating new nations. With the structures of the old system removed, the inhabitants of each country set out on programs to create a postcolonial political, economic, and social order. The obstacles confronting them were myriad and imposing. Indeed, it was only toward 1850, at the end of a 25-year period sometimes known as "the long wait," that the outlines of that new order began to take their definitive form across the region.

POLITICAL MODELS AND THE SEARCH FOR AUTHORITY

Finding foundations of government

One of the most pressing and also most enduring problems that leaders of Latin-American nations faced in the decades after independence was establishing the legitimacy of their new governments. In this regard the break with the colonial system proved traumatic. In Iberian political traditions, power and authority resided to a great extent in the figure of the monarch. Only the monarch had the ability to dominate the church, the military, and other powerful corporate groups in Iberian and colonial Latin-American societies. Representative government and the concept of popular sovereignty, as a corollary, had a weak presence in Iberian political culture. With the Spanish king removed—and with him, the ultimate source of political legitimacy—Creole elites had to find new foundations on which to construct systems of governance that their compatriots would accept and respect.

Although in practice they were unable to abandon the legacies of three centuries of Iberian colonial rule, leaders in Latin America turned generally to other political traditions for solutions to the problem of legitimacy. Adapting models from northern Europe and the United States, they set up republics across the region. Doing so not only helped justify their separation from Spain but also enabled Latin-American elites to try to follow the example of countries they most admired, particularly Great Britain, the United States, and France. Many in the upper classes of Latin-American societies identified political institutions as sources of the economic progress those countries were enjoying. At the same time efforts to implement those political systems in Latin America brought to the region's new countries Enlightenment conceptions of politics based on rationality and a vision of politics as an interaction of individuals who enjoyed specific, definable rights and duties.

Constitutions. Particularly in the first, heady years of independence, elites throughout Latin America exhibited the influence of the Enlightenment in their propensity for producing constitutions. Those documents demonstrated not only attempts to impose rational plans on new nations but also the changing attitudes of elites toward their societies.

The earliest constitutions appeared in Venezuela, Chile, and New Granada in the years 1811–12. The authors of those founding documents rather optimistically intended to create representative government in independent Latin America and to declare inalienable natural rights of liberty, security, property, and equality. To implement those ideas, these constitutions set up a division of power in which the executive was comparatively weak.

Stronger, more centralized governments

From the mid-1810s to mid-century the overwhelming tendency was to move away from those early schemes. With different regions and elite factions battling against each other, the first liberal constitutional governments had failed. Now leaders in the region sought to erect stronger and more highly centralized states, again carefully laying out their programs in constitutions. This shift was not a rejection of foreign models. On the contrary, this change followed the evolution of European political thought; Latin-American elites were now basing their ideas on different foreign theories, turning away from those of Jean-Jacques Rousseau and toward those of more conservative thinkers like Montesquieu and Jeremy Bentham. At the same time, the movement toward stronger executives and more centralized states reflected specific circumstances of these emerging new nations. At first, elites wanted a more powerful state to complete the victory over Spain and then to gain recognition from a Europe by this time dominated by antirepublican attitudes. As political order proved difficult to achieve, many Latin-American leaders also looked to a more centralized state as an instrument against political and civil unrest.

Hopes for a new and stronger government only rarely centred on the idea of monarchy. Leaders in Argentina and Chile discussed the possibility of introducing a constitutional monarchy with a European king at its head. Mexico had emperors, first with Iturbide and then in 1864–67 with the Austrian emperor Francis Joseph's brother Maximilian, and Brazil enjoyed relative stability in a constitutional monarchy that lasted from independence until 1889. Still, such initiatives were temporary and exceptional. Latin Americans encountered a great deal of difficulty in finding suitable European princes to rule their countries. Local figures, furthermore, lacked the necessary authority to be accepted as monarchs. Thus for practical as well as ideological reasons, republics were the rule during the 19th century. As leaders sought greater centralization, they adopted new forms of republicanism. Some, particularly military leaders like Bolívar and generals who had served under him, followed the model of a Napoleonic state. With a powerful president-for-life and a hereditary or life senate, such governments were very close to the structures of constitutional monarchy, although with republican ornamentation. The predominant model, however, was that of the republic that Spanish liberals had set up in 1812. Not all new constitutions after 1815 jettisoned federalism; Mexico in 1824, for instance, embraced that ideal. Overall, Latin America moved toward stronger, more centralized republican governments by the mid-19th century.

Disorder and caudillismo. Written constitutions were not, however, sufficient to enforce order in the new countries of the region. Particularly in the 1825–50 period, Latin America experienced a high degree of political instability. National governments changed hands rapidly in most areas, which only prolonged the weakness and ineffectiveness of the emerging political systems. In Mexico, to take but one example, the years 1825–55 saw 48 turnovers in the national executive. Neither those in power nor those seeking office evinced consistent respect for the often idealistic provisions of constitutions. In some cases the very authors of constitutions broke the rules laid out in them to gain or preserve control over governments. Like any other member of their society, they knew better than to expect their fellow political actors to stay within the strictures of the law. Extralegal maneuvers and the use of force became common elements of politics.

Liberals and conservatives

Much of the conflict that characterized these years consisted of simple disputes over power. Still, by the end of the 1830s and into the 1840s, politics in many areas coalesced around two ideological poles, usually known as liberal and conservative. These groupings were not mass-based political parties in the 20th-century sense but rather factions of the elite; believing the majority of society to be ill-prepared for democracy, both liberals and conservatives intended to construct governments for the people but not by the people. Nonetheless, at times groups of artisans or peasant villages took sides in the factional battles, hoping thus to press their own interests.

The precise definition of the sides in those fights is very difficult, owing to variations between countries and time periods. Urban merchants, rural landowners, and other economic interest groups overlapped so frequently—often within a single family—that it is impossible to generalize about the different origins of political factions. Moreover, the positions taken by one group could be surprising; in Venezuela in the 1840s, for instance, it was conservatives who supported free trade with the exterior, a stance that elsewhere was one of the classic tenets of liberalism. In general, however, one can say that liberals pressed harder for free trade and the rationalization and modernization of their societies—which essentially meant the adoption of European and North American liberal understandings of society as a collection of autonomous individuals. Conservatives, on the other hand, proved more favourable to old institutions, particularly the Roman Catholic church, and to traditional visions of society as grounded in corporate groups. Indeed, in many contexts the question of whether or not to curtail the power of the church was the key point of divergence between otherwise similar liberal and conservative factions.

To an extent, the role that violence or the threat of violence played in politics reflected a militarization brought about by the long period of the wars of independence. Only in Peru and even more so in Mexico did this phenomenon involve the continued influence of a regular, professional military class. Elsewhere the professional military failed to form a coherent interest group, and in many countries civilian politicians managed to control or even reduce the size of their national armies. It was rather in the power of militias and individual military leaders that the militarization of society was most visible. Throughout the region such forces grew to influence or even head national governments.

The military men who rose to positions of dominance were examples of the caudillo, a figure that epitomized this unstable period. Often coming to power through the use of violence, these leaders imposed themselves through the force of their own personalities, their control over armed followers, and their strategic alliances with elite groups. Some caudillos rose to power from humble beginnings, while others came from wealthy, landowning sectors and used their dependent workers as the core of their support. The stereotype of the caudillo as charismatic enough to win the enduring loyalty of his men and skilled enough to ride or fight better than any of them did not, of course, apply to all, but these were domineering and macho leaders. Whatever their social origins, caudillos in the postcolonial period became key political actors, working in alliance with, and at times under the control of, the economically powerful and civilian political leaders of the new nations of Latin America.

In a few cases caudillos contributed to political order. In Chile in the 1830s, for instance, the caudillo Diego Portales was a key figure in the establishment of a comparatively stable government. Allying with conservative elements, Portales helped found a political order that survived his death in 1837. Juan Manuel de Rosas, a caudillo who is said to have been able to outrope and outride his gaucho supporters, imposed a brutal political regime in Argentina from 1829 to 1852. Seeing his homeland split into partisan factions, Rosas sought to ensure a kind of peace by achieving the ultimate victory of one side. His iron-fisted administration, which made use of propaganda and a secret police force, pursued the interests of Rosas and his fellow Buenos Aires ranchers; still, caudillos from other provinces repeatedly tried to oust this violent leader. Indeed, the very foundation of their power in personal relations and in violence meant that the legitimacy of caudillos' rule was always in doubt. Few were able to set up networks of alliances that could withstand the challenges of new leaders who emerged with their own armed supporters and wealthy allies. The system of caudillismo was a volatile one. Although the general type continued to exist throughout the 19th century, it was the postindependence period that represented the golden age of the caudillos.

Juan Manuel de Rosas

ECONOMIC OBSTACLES

Complicating the construction of stable, constitutional governments in the decades after independence were the economic circumstances that prevailed in the period. Creoles who had expected the dismantling of colonial restraints on Latin-American economies to produce a wave of new wealth found their hopes dashed in the 1820s. In many ways the region's economies were poorer and less integrated in the first decades after independence than they had been in the late colonial period. Political disorder was both a cause and result of this situation. Unable to rely on old taxes for revenue, new governments commonly found themselves in tight financial straits. Their resulting weakness contributed to political instability, which at the same time impeded the reorganization of economic systems.

The effects of war

The wars of independence contributed to the disappointing postwar economic picture. In some areas, such as Venezuela, damage from the wars was extensive. Even where the destruction of human life and economic resources was less widespread, disruptions in financial arrangements and systems of labour relations provoked a decline in important economic sectors. Mining suffered particularly in many countries. The richest mineral producer, Mexico, needed fully half a century to regain its preindependence levels of production.

As they emerged from their battles for emancipation, the new nations encountered other difficulties. The mere fact of political independence did not eliminate long-standing problems of transportation, but it did break down some traditional commercial networks. The entrance of foreign merchants and imported goods, although on a much more limited scale than would later be the case, led to competition with, and in some areas the displacement of, local traders and producers. Apart from loans that left several countries in debt, the region received little capital from foreign sources. The departure of, or discrimination against, peninsular Spaniards drastically reduced what had been a major source of skilled labour and administrative know-how, as well as capital for investment. Relatively few exports, such as coffee, sugar, and cattle products, found world markets favourable enough to stimulate the expansion of their production in Latin America. Colonial patterns had been destroyed, but the economies of the region had not yet found a consistent new orientation.

SOCIAL CHANGE

Mobility and hierarchy. The Creole elites who had headed the independence cause throughout Latin America had no intention of losing their social, economic, and political power in the construction of new nations. Managing to solidify and even expand their influence after the removal of colonial administration, these elites emerged as the great beneficiaries of independence.

The situation of other social groups and institutions was more mixed. Leaders across the region quickly eliminated the system of separate ethnic castes. Persons of mixed race were, in theory, to have the same legal rights as members of the white upper classes. Indeed, the period of independence saw the ascension of individual mestizos and castas to positions of prominence. Service in the wars was particularly useful in this regard. Men such as the mulattoes Manuel Piar and José Padilla in Venezuela and New Granada, respectively, rose to the rank of general in Bolívar's armies. In practice, however, the old hierarchies did not fall so easily and continued on informally. Those nonwhites who managed to achieve the status of elites were clearly exceptions to the general rule. The destruction of the caste system allowed for only limited loosening of racial and class hierarchies. Indeed, both Piar and Padilla were executed under rather questionable circumstances.

The position of Indians changed rather slowly in the postindependence era, despite some early and energetic initiatives. Spain had ended Indian tribute in 1810, and in the years after that several Latin-American nations saw fit to repeat that measure with abolitions of their own. More generally, leaders frequently spoke of breaking down the barriers between the indigenous and more Hispanized sectors of their societies. Still, in the aftermath of independence, governments tended to reverse their positions toward Amerindian populations. The countries of the Andes, for example, reinstated Indian tribute, albeit under different names. Bolivian governments derived as much as 80 percent of their revenues from that source through mid-century. Full-scale attacks on indigenous communities waited until later in the century.

Slavery

Strong measures against African slavery similarly appeared in many areas by the late 1820s. Lawmakers declared the children of slaves to be free, banned the slave trade, or even ended slavery itself. Once again, however, there was a pattern of backsliding, so that, where slave labour played a significant economic role, the final abolition of the institution of slavery came about in most countries only about 1850. The growth of sugar production in Cuba and coffee production in Brazil, furthermore, meant that those two slave societies continued to flourish. Both areas continued to receive large numbers of new enslaved workers from Africa until after mid-century (1865 in Cuba, 1851 in Brazil) and only abolished slavery in the 1880s (1886 in Cuba, 1888 in Brazil).

Social institutions. Both as part of their ideological commitment to liberal individualism and as a means of increasing the power of their new states, leaders in the postindependence years tried to establish their control over the formidable colonial institutions of the Roman Catholic church and the military. Success came more easily in the case of the military. Only in Mexico and to a lesser extent in Peru did professional armies form fairly coherent interest groups pressing for the maintenance of their traditional privileges. After mid-century, however, those special privileges were lost even in these countries. The church, on the other hand, though losing a great deal of power, held on to a position of influence in much of the region. Armies of independence and some subsequent governments took over church properties and resources to meet their financial needs. In Buenos Aires and Montevideo, liberals were also able to trim the privileges of the church; elsewhere, however, attempts to do so either appeared later or, as in Mexico and Guatemala, provoked serious conflicts.

The new order, 1850–1910

POLITICAL AND ECONOMIC TRANSITIONS, 1850–70

The first decades of the second half of the 19th century represented the beginnings of a fundamental shift in the still-young nations of Latin America. At the heart of this transition was a growing orientation of the economies of the region to world markets. As Europe and North America experienced a second wave of industrialization, they began to reevaluate the economic potential of Latin America; the region looked to them increasingly like a vital source of raw materials for the expanding economies of the North Atlantic. To take advantage of the possibilities that this conjuncture opened, elites in Latin America directed their countries ever more toward export economies. That change also entailed a series of social and political developments that, especially from the 1870s on, constituted a new order in Latin America. The 1850s and '60s were merely a transitional period, however, as political conflicts and civil wars broke out in Mexico, Venezuela, and elsewhere, postponing the consolidation of the general shift.

THE LIBERAL OLIGARCHIC AGE, 1870–1910

Neo-colonial order

The order that took shape in the last decades of the 19th century is often called neocolonial, as a way of suggesting that the internal and external structures characterizing the region maintained overall similarities to those of the period of Iberian colonial rule. To a great extent this is a useful description. As in the colonial period, the region was tremendously vulnerable to outside events and foreign nations. Although many Latin-American elites profited from the new order, they ceded a degree of control over their countries to the industrializing economies of the North Atlantic. For much of the 19th century Britain was the predominant power in the region, followed by the United States, France, and Germany. By the end of the 1870–1910 period the United States managed to supplant Britain. As in colonial times, Latin America continued to be largely an exporter of raw materials and an importer of manufactures. Furthermore, despite some legal changes, social relations had not undergone revolutionary change. Broad hierarchies of race and class continued to define social relations. In the countryside in particular the figure of the *patrón* (boss or patron) maintained dominance over both physical resources and persons of lower status. The role of such men as patriarchs in their households demonstrates further that the relative positions of men and women had not become noticeably more equal; although not accepted by all, definitions of women as weaker than men and fit primarily for domesticity were still the norm.

The patterns of 1870–1910 were not, however, mere copies or repetitions of colonial trends. Along with the similarities to earlier conditions came profound economic, social, and political changes. In this regard the term "neocolonial" does not capture the complexity and dynamism of this period in Latin-American history.

Export economies. Through the mid-19th century many interests in Latin America had doubts about the wisdom of opening their economies to the world. In countries like Peru and Colombia, artisans and other producers, as well as some merchants, persuaded their governments to set up barriers against the entrance of foreign competition. By the 1860s and '70s, however, such protectionism was swept away by a wave of free-trade liberalism. Domestic production of textiles and other goods proved incapable of doing more than merely surviving. When the great impulses toward direct links to Europe and the United States emerged, elites across Latin America turned their backs on the artisans and weavers in their countries and enthusiastically welcomed in manufactures from England, the United States, and other nations. The doctrines of liberalism—from free trade internationally to open markets domestically—became hegemonic.

Besides the upsurge in international demand for Latin-American primary goods, the factors fueling the rise of export economies included foreign investment and technological innovations brought from the industrializing countries. A wide range of products were affected by the increase in demand, from consumer goods such as sugar, coffee, wheat, and beef to industrial products like rubber and minerals. Old products such as silver recovered and surpassed earlier levels of production, while other new products appeared. One spectacularly successful new export from mid-century to the 1870s was guano, or seabird dung, which was mined on the islands off the Peruvian coast and sold to Europe as a fertilizer. When new chemical fertilizers shut down foreign markets for guano, nitrates and copper from the arid regions of northern Chile entered the scene as profitable new mining products for export.

The lack of capital that had plagued Latin America in the immediate postindependence period was resolved now by injections of foreign capital on a scale previously unknown. Investments from Europe provided much of the financial support for infrastructural improvements. British and other foreign firms constructed railways, streetcar systems, and electric networks, often getting guarantees of profits on their investments and other favourable concessions from local authorities. At the same time, some ominous signs appeared; often borrowing against projected export earnings, the Peruvian and other governments ran up large foreign debts in the late 19th century.

Technological advances

Along with financial capital came technology, in such forms as barbed-wire fencing, refrigeration, steam engines, and mining equipment. With access to credit, both foreign and domestic producers were now able to adopt such technologies, thereby increasing the size and efficiency of their production for export markets. The Cuban sugar economy, for example, underwent major changes linked to the creation of highly capitalized central mills that used new processing machinery to increase refining capacity and benefited from new transportation technology to ease the sale to export markets. Indeed, perhaps the single most important technological advance was the railroad; in this bold age of construction, railroads thrust out across much of Latin America, speeding transportation between productive zones and urban centres and ports. The spread of rail lines brought year-round transportation to regions that had lacked it. Moreover, by reducing freight costs, railways fostered the production of bulk commodities like beef and coffee. Together with the introduction of steamship lines in the Magdalena, Orinoco, La Plata–Paraná, and other river systems, the railroad thus opened up the possibilities for exports of primary goods. Communications also improved with the introduction of telegraph lines, which by the 1870s linked parts of Latin America directly with Europe. Both the new investments and technology transfers served to facilitate production and export of the primary goods that industrializing economies sought. Latin America underwent a thorough integration into the world economy.

Even as it opened up areas of lucrative production, this new orientation of Latin-American economies imposed certain limits. The concentration on exports of primary goods and the competition of imported manufactures with domestic products served as powerful disincentives to economic diversification. Some areas, like Cuba with sugar and Central America with coffee, fell into patterns of monoculture, in which an entire national economy was dependent on the health of one particular crop. Even where more than one product was central to a country, the reliance on these exports made Latin-American economies vulnerable to shifts in demand and prices on the world market, as well as to local conditions influencing production.

Although the new order favoured a focus on raw materials production, some areas experienced the beginnings of industrialization. Particularly in capitals that served as commercial as well as administrative centres, such as Buenos Aires, the late 19th and early 20th centuries witnessed the rise of tertiary sectors as well. The increased volume of production and trade spawned a wide range of services that created jobs in manual labour in docks and processing plants and white-collar work in both the government and private firms. Manufacturing sprang up in countries like Chile and Brazil, often starting with the production of cheap textiles and other relatively simple goods that could compete with low-end imports. Some of the financing for such ventures came from abroad. A significant and often underestimated portion of the capital that the new systems of banking and finance provided for early manufacturing efforts, however, consisted of local capital. Groups that had grown wealthy and powerful in the export economy began to diversify into manufacturing in areas like São Paulo. Still, the transition from exporters of primary goods to producers of manufactures was a difficult one in which the region participated unevenly. Most notably in Central America and in the Caribbean, the local elites' activities were largely restricted to the production of primary export goods, and economies retained more of a neocolonial orientation.

Capitalism and social transitions. The social ramifications of the rise of export economies were vast. The acceleration of the export economies and related commerce fostered a tendency toward urbanization. The period was one of general population growth in much of Latin America, most spectacularly in the temperate, staple-producing zones of South America. Within the overall increase, the rise of cities was particularly noteworthy. More than simple size was involved; cities like Rio de Janeiro, Buenos Aires, and Mexico City became sophisticated, cosmopolitan urban centres. Urban reforms, many inspired by the sweeping transformation of the French capital under Napoleon III and his city planner, Georges-Eugène Haussmann, allowed cities to vie with each other for the title of "Paris of South America." At the same time, incipient industrialization brought conflicts between urban workers and capitalists. Workers had for decades been organizing themselves into mutual aid societies and other nonideological associations. At the end of the 19th century and the beginning of the 20th, new groups began to emerge. At times with the special participation of recent European immigrants, workers established trade unions, pressing their interests with strikes and other activities. In this early phase, ideologies of anarchism and anarcho-syndicalism had particular influence in many areas. By the early 20th century, moreover, the growth of government and service sectors created urban middle classes that were ready to enter politics.

In the countryside, social relations underwent greater change over a short period than at any time since the conquest. Increasing ties to the capitalist world economy did not always lead to wage labour but rather fed the diversification of work relations. In fact, one of the predominant tendencies of the period was the strengthening or even extension of certain nonwage forms of labour. In parts of Peru, Mexico, Central America, and other areas, debt peonage became key to export agriculture. In this system employers or labour agents advanced a certain sum to workers, who would then have to labour on a ranch or plantation to pay off their debt. Because of manipulations by the owners, however, the workers often found that their indebtedness only grew the longer they toiled. In some places debt peonage amounted to a form of de facto slavery. Countries like Argentina and Guatemala, moreover, saw the rise of so-called vagrancy laws, by which authorities could force unattached gauchos or peasants to work on large rural estates. In the Central Valley of Chile existing tenancy arrangements suffered modifications that cut back the rights and privileges of poor rural workers. Brazil and Argentina, on the other hand, experienced the emergence of unique systems of farming by European immigrants, which brought modern wage systems to important areas of their economies. Indeed, in those countries, immigration of Italians, Spaniards, and other Europeans transformed the ethnic composition and habits of whole regions. Argentina alone received almost 2.5 million people in this period.

Diversification of work relations

Throughout Latin America the position of rural workers came under attack from the large plantations, ranches, and estates that were expanding to take advantage of potential profits from the export economies. In south-central Brazil coffee plantations spread westward, pushing back small foodstuff production; in Argentina the ranching frontier pressed southward, displacing indigenous groups. Peasants and indigenous communities had resisted encroachment by neighbouring estates throughout the early national period and continued to do so into the 20th century. Nevertheless, the balance of power was shifting in favour of the large landowners. Early liberal moves to break up communal land holding paled beside the more energetic initiatives of the later 19th century. Although indigenous communities survived in the Andes, Mexico, and Central America, they commonly lost land, access to water and other resources, and some of the limited autonomy they had enjoyed.

The Roman Catholic church also was the target of ever more aggressive liberal attacks after mid-century. As in the case of indigenous communities, the justification for those assaults was based in liberal ideology; politicians argued that property had to be placed into the hands of individuals because they would be more likely to develop it efficiently and thus contribute to economic progress. In much of Latin America the church had been the preeminent source of capital and a major property owner. When governments began large-scale appropriations of church holdings, it was thus a major occurrence. In Mexico, where the church controlled as much as a third of all rural real estate, such government efforts met repeated, violent resistance. By the 1870s, however, the takeover of church properties was irreversible.

Oligarchies in power. Along with the export economies came political transitions. The increased revenues that burgeoning commerce provided allowed elites to consolidate more orderly political systems in some countries. Political unrest continued, however, in others; Colombia, for instance, experienced a series of civil wars toward the end of the 19th century.

Across the region groups tied to the export economies came to dominate politics in this era. In 1871 Guatemalan liberals linked to the rising coffee sector ousted the conservative regime that had controlled the country since 1838. The years 1876–1911 in Mexico, meanwhile, marked the iron-fisted rule of Porfirio Díaz, who began his career as a liberal fighting under a banner of election for one term only and ended up as a dictator who customarily manipulated his country's political structures to ensure that he and his allies would remain in power. That regime, known as the Porfiriato, was a particularly clear example of the late 19th-century regimes' ties to the new economic order. The Díaz government, like other progressive dictatorships in Latin America, worked to promote railroad construction, to force reluctant peasants and indigenous groups to work on rural estates, to repress popular organizing, and in other ways to benefit the elites of that order. Through such initiatives the governments of the day diverged from pure liberal tenets according to which the market alone determines the shape and nature of economic change. In many countries ruling groups began to adopt the ideas of positivism, and ideology stressing a scientific analysis of human history and efforts to accelerate progress. In Brazil the decentralized old republic, dominated by rural elites, replaced constitutional monarchy in 1889 and took as its motto the positivist slogan "Ordem e Progresso" ("Order and Progress"). That phrase summed up what the ruling groups in Brazil and across Latin America sought in the mature age of export-oriented transformation—the maintenance of the hierarchies that they dominated and the achievement of prosperity and a "civilization" that represented an approximation of North Atlantic models. Thus both oligarchic republics and liberal dictatorships evolved as part of the new order of the 1870–1910 period.

Porfiriato

(R.A.Ki.)

New order emerging, 1910–45

The advances in economic growth and political stabilization that were evident in most of Latin America by the early 20th century came up against an array of challenges as the century wore on. The forward momentum was not necessarily lost—although Mexico experienced negative economic growth along with great political turmoil during the first decade of the Mexican Revolution beginning in 1910—but some partial changes of direction occurred, and new problems kept emerging. The challenges were of both internal and external origin, ranging from steady population increase for the region as a whole to the consequences of Latin America's ever-closer incorporation into the world economy.

ECONOMIC AND SOCIAL DEVELOPMENTS

World war and world trade. Few Latin Americans felt strong emotional identification with either of the contending alliances in World War I (1914–18), except for the immigrant communities in southern South America and the ranks of generally Francophile liberal intellectuals. Of the major countries, only Brazil followed the example of the United States in declaring war on Germany, while Mexico and Argentina, which respectively saw the United States as a bullying neighbour and a hemispheric rival, vied for a leadership role in behalf of Latin-American neutrality. Yet all countries were affected by the wartime disruption of trade and capital flows, particularly those that had in recent years successfully penetrated European markets with their own exports and become important consumers of European goods and financial services. Argentina was an obvious example. The outbreak of war brought a sharp decline in its trade as the Allied powers diverted shipping elsewhere and Germany became inaccessible. Although exports soon recovered, mainly in the form of meat to feed Allied troops, imported manufactures were scarce because overseas factories were devoted to war production, and scarcity drove up prices.

Booms and busts

Wartime disruptions were only temporary, and they gave way to a frenzied boom in the immediate postwar period as Latin-American exporters cashed in on pent-up demand in the former warring powers. An extreme case was the "dance of the millions" in Cuba, where the price of sugar reached a peak of 23 cents per pound in 1920, only to fall back to 3.5 cents within the space of a few months, as European production of beet sugar returned to normal. Similar postwar booms and busts occurred elsewhere, even if less sharply, and demonstrated some of the hazards of Latin America's increasing dependence on the world economy. Those hazards were underscored again by the costly program Brazil felt compelled to undertake to support the price of coffee, buying up surplus production and keeping it off the market. First tried in 1906 and briefly repeated during the war, this "valorization" policy was reinstated during the 1920s in the face of persistent weakness of the world coffee price. Yet one reason for the latter was the expansion of cultivation in other Latin-American countries, above all Colombia, which by the end of World War I had emerged as the second leading producer—encouraged by, among other things, the Brazilian price support efforts.

Conditions in the world market were in the last analysis unfavourable for Latin America's terms of trade, since demand for most of the primary commodities that the region specialized in was not keeping pace with the growth of production. Nevertheless, the decade of the 1920s was generally a period of economic growth and renewed optimism. All countries continued to pursue an outward-directed growth strategy insofar as they pursued a conscious strategy at all, placing few impediments in the way of import-export trade. Foreign investment also resumed on a massive scale and now came chiefly from the United States. New capital flowed both into productive activities, like the Venezuelan petroleum industry (controlled by U.S., British, and Dutch interests and by the late 1920s the world's leading exporter though not producer), and into loans made by Wall Street bankers to Latin-American governments.

The emerging force of nationalism. The growing importance of foreign capital inevitably provoked a nationalist backlash, which reinforced the cultural nationalism already strong among groups of intellectuals and the anti-imperialist sentiment provoked by U.S. intervention around the Caribbean and in Mexico. Cultural nationalism was associated above all with conservatives who cherished the Iberian heritage as a shield against corrupting Anglo-Saxon influences, while the leading anti-imperialist spokesmen tended to be leftist. Incipient left-wing parties and labour unions were also in the forefront of economic nationalism, because, among other reasons, foreign-owned firms provided a more popular target than local enterprises. British nitrate investors in Chile thus faced serious labour unrest, as did the Boston-based United Fruit Company, hit by a violent strike in late 1928 in the Colombian banana zone. Petroleum investors in Mexico faced serious labour unrest in addition to a simmering conflict with the government itself over the control of subsoil resources, which the new constitution of 1917 had declared exclusive property of the nation.

Effects of the depression

A further escalation of economic nationalism came with the world economic depression of 1929 and after, though more as a defensive reaction than as a conscious policy. For Latin America, the depression put an abrupt end to the inflow of foreign capital and at the same time brought a drastic decline in the price of the region's exports, which in turn reduced the capacity to import and the governments' revenues from customs duties. At one point, a pound of Cuban sugar was selling for less than the U.S. tariff on the sugar. In response to the crisis, Latin-American countries raised their own tariffs and imposed other restrictions on foreign trade. Even if the immediate purpose was conservation of scarce foreign exchange rather than the theoretical goal of increasing economic independence, the result was a decided impetus to domestic manufacturing, whose beneficiaries later appealed to nationalist sentiments to preserve the gains made. In Colombia, textile production increased during the 1930s at a faster rate than in England during the Industrial Revolution, despite the fact that the government continued to see protection of the coffee industry as its primary economic mission. But manufacturing made important gains in almost all the larger Latin-American nations, which already before the depression had begun the development of an industrial base. It remains to be said, however, that, except for Mexico with its well-established iron and steel industry, manufacturing still consisted almost wholly of consumer goods production.

On another front, to save available jobs for native inhabitants, numerous countries adopted measures during the depression that required a given percentage of a company's employees to be citizens. In Brazil, for similar reasons, tight restrictions were imposed on the flow of immigrants. Even without restrictions, however, and despite the fact that some countries recovered quickly from the effects of the depression, Latin America in the 1930s was simply not as attractive to immigrants as before.

Population and social change. In some countries the life of most inhabitants seemed little changed in 1945, at the end of World War II, from what it had been in 1910. This was the case in Paraguay, still overwhelmingly rural and isolated, and Honduras, except for its coastal banana enclave. Even in Brazil, the *sertão,* or semiarid backcountry, was barely affected by changes in the coastal cities or in the fast-growing industrial complex of São Paulo. But in Latin America as a whole more people were becoming linked to the national and world economies, introduced to rudimentary public education, and exposed to emerging mass media.

Even in Argentina, Brazil, and Cuba, where the number of immigrants had been significant up to the depression—in Cuba's case above all from Spain—population growth was mainly from natural increase. It was still not explosive, for, while birth rates in most countries remained high, death rates had not yet been sharply reduced by advances in public health. But it was steady, the total Latin-American population rising from roughly 60 million in 1900 to 155 million at mid-century. The urban proportion had reached about 40 percent, though with great differences among countries. The Argentine population was approximately half urban by the eve of World War I, fewer hands being required to produce the nation's wealth in the countryside than to process it in the cities and provide other essential urban services. In the Andean countries and Central America, however, urban dwellers were a decided minority even at the end of World War II. Moreover, the usual pattern was that of a single primate city vastly overshadowing lesser urban centres. In Uruguay in the early 1940s, Montevideo alone had 800,000 inhabitants, or over one-third of the nation's total, while its closest rival contained about 50,000. Yet even that was as many as lived in Tegucigalpa, the capital of Honduras.

The working and middle sectors

Latin America's population is less easy to classify in terms of social composition. Rural workers still made up the largest single group, but those loosely referred to as "peasants" could be anything from *minifundistas,* or independent owners of small private parcels, to seasonal hired hands of large plantations; with different degrees of autonomy and different linkages to national and world markets, they were far from a cohesive social sector. What such rural workers most clearly had in common was grossly inadequate access to health and education services and a low material standard of living. A socioeconomic and cultural gulf separated them from traditional large landowners as well as from the owners or managers of commercial agribusinesses.

In the cities an industrial working class was more and more in evidence, at least in the larger countries, where the size of the internal market made industrialization feasible even with low average purchasing power. However, factory workers did not necessarily form the most important sector, to some extent because the growth of cities had been more rapid than that of the manufacturing industry. São Paulo in Brazil and Monterrey in Mexico won fame chiefly as centres of industry, but more typical was the case of Montevideo, a commercial and administrative centre first and foremost that attracted the lion's share of the country's industry because of its preexisting leadership in population and services rather than the other way around. Moreover, port, transportation, and service workers—or miners, as in the Chilean nitrate fields—rather than factory workers usually led the way in union organization and strike actions. One reason was the high proportion of women workers in early factories, who, though even more exploited than male workers, were perceived by radical activists as less-promising recruits than stevedores or locomotive firemen.

In urban settings the most important social development in the short run was the steady expansion of middling white-collar and professional groups. The extent to which these can be termed a "middle class" is open to question, for, while "middle" by the economic indicators of property and income, they were often ambivalent about their place in society—uncertain whether to embrace the work and savings ethic conventionally associated with the middle class of the Western world (or later Japan) or to try to emulate traditional elites. The middle sectors were, in any case, the chief beneficiaries of the expansion of educational facilities, which they strongly supported and used as means of upward mobility. Urban workers, for their part, had access to primary education but rarely secondary and higher; at least they were now mainly literate, whereas most rural Latin Americans still were not.

Lack of formal education had long reinforced the relative isolation of the peasantry from political currents at their nations' centres, not to mention from new fads and notions from abroad. Yet, starting in the 1920s, the rapid spread of the new medium of radio throughout Latin America exposed even illiterate people to an emerging mass culture. Additions to transportation infrastructure also contributed to greater integration of isolated population clusters. The most essential rail lines had already taken shape by 1910, but the coming of automotive transport led to a major upgrading and extension of highways, and the airplane introduced an entirely new mode of transportation. One of the oldest airlines in the world is Colombia's Avianca, whose founding (under a different name) in 1919 was of particular importance for a country where railroad and highway building had lagged because of difficult topography. Air travel similarly played a key role in knitting together far-flung sections of Brazil previously connected by coastal steamer. Transport improvements of all kinds favoured the creation not only of national markets but of shared national cultures, in the latter respect reinforcing the effects of popular education and radio.

CHALLENGES TO THE POLITICAL ORDER

The economic and social changes taking place in Latin America inevitably triggered demands for political change as well; political change in turn affected the course of socioeconomic development. As the 20th century opened, the most prevalent regime types were military dictatorship—exemplified by that of Porfirio Díaz in Mexico and after 1908 Juan Vicente Gómez in Venezuela—and civilian oligarchy—as in Chile, Argentina, Brazil, or Colombia. Even in Díaz's Mexico the constitution was not entirely meaningless, while civilian governments commonly used some combination of electoral manipulation and restricted suffrage to keep control in the hands of a small minority of political leaders allied with landed and commercial elites. Neither dictatorial nor oligarchic regimes gave due representation to the majority of inhabitants.

The Mexican Revolution. The immediate challenge to existing regimes in country after country came from disaffected members of the traditional ruling groups and from the expanding middle sectors resentful of their exclusion from a fair share of power and privilege. This was evident at the outset of Latin America's bloodiest 20th-century civil conflict, the Mexican Revolution of 1910, when a dissident member of the large landowning class, Francisco Madero, challenged Díaz for reelection, lost, and rose in rebellion, promising to bring genuine political democracy to Mexico. The dictatorship, decaying from within, collapsed, but it was many years before the country settled down, since Madero's uprising unleashed forces that neither he nor anyone else could control. Miners, urban workers, and peasants saw an opportunity to seek redress of their own grievances, while rival revolutionaries bitterly fought against each other. The end result was a system built around an all-powerful political party—the Institutional Revolutionary Party (Partido Revolucionario Institucional; PRI), as it ultimately called itself—that skillfully co-opted labour and peasant organizations. More benefits accrued to labour leaders than to the rank and file, and implementation of the land reform proclaimed by the new constitution of 1917 was mostly halfhearted until the presidency of Lázaro Cárdenas (1934–40). But it appeared that almost everybody received something, and

Francisco Madero

after Cárdenas Mexico became a model of political stability in Latin America.

Broadening of political participation. The Mexican Revolution evoked widespread admiration elsewhere in Latin America, especially for its commitment to socioeconomic reform, but the Mexican political system had few imitators. In the Southern Cone, a common pattern was the broadening of participation within a more conventional democratic system where at least the middle sectors gained a meaningful share of power and benefits. This happened in Argentina following an electoral reform of 1912 that made universal male suffrage effective for the first time and paved the way for the Radical Civic Union party, with strong middle-class support, to take power four years later. In Chile a reformist coalition won the election of 1920, but strife between president and parliament brought a relapse into instability and short-lived military dictatorship. By the time Chile returned to stable political life in 1932, it had been equipped with a new constitution that was less susceptible to oligarchic obstructionism and an apparatus of social legislation that benefited both the middle class and urban workers, though it largely ignored the peasantry. However, Uruguay outstripped all others both in political democratization and as a pioneer welfare state, with minimum-wage legislation, an advanced social security system, and much else, even before 1930.

Elsewhere the record was mixed. Costa Rica came close to approximating the pattern of the Southern Cone, and in Colombia the Liberal Party, after its return to power in 1930, went partway toward incorporating labour as an actor on the national scene. Ecuador in 1929 became the first Latin-American nation to adopt woman suffrage, though it still required literacy to vote (and far fewer women than men could read). Within four years Brazil, Uruguay, and Cuba—of which only the first retained a similar literacy test—had followed suit. But in Peru a president who flirted too far with social and political reform at the time of World War I was ousted by military coup. In the following decade the banner of reformism in Peru was taken up by Víctor Raúl Haya de la Torre, founder of the Aprista party and heavily influenced by the example of the Mexican Revolution. The Apristas' program combined economic nationalism with Latin-American solidarity and called for incorporation of the Indians into the mainstream of national life, but the party never gained control of government until the 1980s, by which time it had lost much of its original character. In Venezuela, thanks to oil revenue and effective use of the military, Juan Vicente Gómez stayed firmly in control as dictator until his final illness in 1935; and in Brazil the oligarchic regime of the so-called Old Republic held on until the economic crisis of the Great Depression through careful sharing of power among political factions of the largest states.

The Aprista party

Expanding role of the state. The world depression—which saw governments changed by irregular means in every Latin-American country except Colombia, Venezuela, Costa Rica, and Honduras—temporarily ended the progress being made toward political democracy. Even where constitutional rule was not interrupted, chief executives felt the need (as also in the United States) to take emergency measures, and the enlargement of government functions in dealing with the economy outlasted the emergency itself. At the same time, leaders everywhere were coming to the conclusion that social ills must be ameliorated, if only to ward off revolutionary threats from below. Various countries (such as Colombia in 1936 and Cuba in 1940) adopted constitutional reforms incorporating the principle already enshrined in Mexico's constitution of 1917, of subordinating property rights to social need.

Brazil had actually pioneered large-scale state intervention in the economy with its coffee "valorization" program, which was finally abandoned during the depression as too expensive; but between 1930 and 1945, under President Getúlio Vargas, the national government for the first time actively sponsored social legislation, encouraged labour unions while tying them closely to the state, and began construction of a major iron and steel complex under state auspices. Vargas was an authoritarian ruler but a constructive one. Nor was he the only military or civilian strongman who moved to expand the functions of the state both to take the edge off worker discontent and, if possible, to strengthen the national economy against new emergencies. A paradoxical but instructive example was Cuba's notoriously corrupt Fulgencio Batista, who in 1933 staged a military coup to overthrow a government of the reformist Authentic Party, then preserved most of its social and labour reforms and added some more. After sponsoring the liberal Cuban constitution of 1940, he managed to become a democratically elected president.

Socialism, Communism, Fascism. Latin America in the first half of the 20th century was feeling the impact of outside events not only on its economy but also politically, by the spread of imported ideologies and through the examples both of President Franklin D. Roosevelt's New Deal in the United States and of emerging totalitarianisms of the left and right in Europe. The European anarcho-syndicalism that had provided a model for many of Latin America's earliest radical cadres declined sharply in importance after World War I. Henceforth, the left consisted of socialist parties of generally moderate bent, inspired in large part by European social democracy; breakaway socialists who admired the Russian Revolution of 1917 and proceeded to found communist parties in their own countries; and, not least, such strictly Latin-American expressions as the Mexican agrarian reform movement. Socialist parties were strongest in the Southern Cone, the Chilean briefly gaining a share of national power as a member of a Popular Front government elected in 1938. The communists were also strong in Chile but first entered a national administration in Cuba, after Batista had been elected president with their support in 1940. Once the Soviet Union entered World War II in 1941, communist parties in several other countries, including Brazil and Nicaragua, formed alliances with local strongmen, but they nowhere became a true mass party, and an exaggerated fear of Bolshevism on the part of Latin-American elites meant that the communist parties were subject to widespread repression except during the war itself.

Parties of the left

Some other political organizations were frankly influenced by European fascism, but in most countries their membership was numerically insignificant. The chief exception was Brazil, whose green-shirted Integralistas (Ação Integralista Brasileira) emerged as the largest single national party in the mid-1930s until involvement in a foolhardy coup attempt led to their suppression. Hence the influence of fascism was more often exercised through homegrown authoritarians who were attracted to certain aspects of it but carefully avoided any open embrace. Vargas was one such leader, who, after suppressing the Integralistas, put the finishing touches on his own dictatorial regime, officially dubbed Estado Novo or "New State."

Good Neighbor Policy and World War II. One reason Latin-American nations avoided an overly close association with fascism was a desire not to offend the dominant power of the hemisphere, the United States. During the 1920s it had already begun a retreat from the policy of active intervention in Latin America. This policy, adopted in the aftermath of the Spanish-American War and the United States' open support of Panamanian secession from Colombia, had featured the creation of formal and informal protectorates over many Caribbean and Central American states. Franklin D. Roosevelt completed the shift. His domestic policies were much admired in Latin America and in some cases copied by moderate reformists, but his Good Neighbor Policy won the warm approval of almost all Latin-American rulers, since it entailed formal renunciation of the right of intervention in favour of peaceful cajoling and assorted economic, military, and technical aid programs. These programs were launched on the eve of World War II to help hemispheric neighbours prepare for the emergency. They were expanded after the start of the conflict, whose economic impact on Latin America was generally comparable to that of World War I but more intense because of the earlier and deeper involvement of the United States. The war emergency naturally gave still further impetus to the development of national industries to replace scarce imports.

The Good Neighbor approach proved far more effective in promoting U.S. hegemony than the occasional dispatch of gunboats. In 1938 Roosevelt calmly accepted Mexico's expropriation of the petroleum installations of U.S. and British companies, and he was rewarded several times over when Mexico loyally cooperated with the United States in World War II, even sending an air force squadron to serve in the Philippines. The one other Latin-American country to send forces overseas was Brazil, which put an expeditionary force in Italy. In general Latin America's wartime collaboration left little to be desired. In the end all countries not only broke relations with the Axis Powers but declared war, though Argentina took the latter step only at the last possible moment, in March 1945.

Latin America since the mid-20th century

THE POSTWAR WORLD, 1945–80

In Latin America as elsewhere, the close of World War II was accompanied by expectations, only partly fulfilled, of steady economic development and democratic consolidation. Economies grew, but at a slower rate than in most of Europe or East Asia, so that Latin America's relative share of world production and trade declined and the gap in personal income per capita separating it from the leading industrial democracies increased. Popular education also increased, as did exposure to the mass media and mass culture—which in light of the economic lag served to feed dissatisfaction. Military dictatorships and Marxist revolution were among the solutions put forward, but none were truly successful.

Economic agenda and patterns of growth. The economic shocks delivered by the depression and two world wars, in combination with the strength of nationalism, tilted economic policy after 1945 strongly toward internal development as against the outward orientation that had predominated since independence. The outward policy had been partially undermined by the trade controls and industrial promotion schemes adopted essentially as defensive measures in the aftermath of the depression and during World War II. Now, however, a reorientation of policy was explicitly called for by some of Latin America's most influential figures, such as the Argentine economist Raúl Prebisch, head of the United Nations Economic Commission for Latin America. Prebisch and his followers insisted that the terms of trade and investment in the contemporary world were stacked in favour of the developed industrial nations of the "centre" as against the developing nations of the "periphery." Their strategy therefore included emphasis on economic diversification and import substitution industrialization (ISI) for the sake of greater economic autonomy. They called for economic integration among the Latin-American countries themselves, with a view to attaining economies of scale. And they recommended internal structural reforms to improve the economic performance of their countries, including land reform both to eliminate underutilized latifundios and to lessen the stark inequality of income distribution that was an obstacle to growth of the domestic market.

Import substitution industrialization (ISI)

In the small Caribbean and Central American republics and also some of the smaller and poorer South American nations, the prospects for ISI were sorely limited by market size and other constraints, and governments still hesitated to promote manufacturing at the expense of traditional primary commodities. But in countries accounting for a disproportionate share of Latin America's population and gross domestic product (GDP), the new approach received full play through protective tariffs, subsidies, and official preferences. Overvalued exchange rates, which hurt traditional exports, made it easier to import industrial machinery and equipment. Manufacturing costs generally remained high, and factories were overly dependent on imported inputs of all kinds (including foreign capital), but advances were not limited to consumer goods production. In all major countries the output of intermediate and capital goods rose appreciably too. In both Argentina and Colombia the state undertook construction of a steel industry, and in numerous other ways national governments further expanded their economic role. Brazil nationalized its incipient oil industry in 1953, creating the state firm Petrobrás that eventually ranked alongside Mexico's PEMEX (outcome of the 1938 oil expropriation) and Venezuela's PETROVEN (1975) as one of Latin America's three largest economic enterprises, all state-run.

Starting in 1960 with agreements fostering economic union, such as the Latin American Free Trade Association and Central American Common Market, and continuing with the Andean Pact of 1969, some progress was made toward regional economic integration, but the commitment to eliminate trade barriers was not as strong as in postwar Europe. Intra-Latin-American trade increased, but probably not much more than would have happened without special agreements. In any case, quantitative economic growth was visible almost everywhere. It was evident even when expressed as per capita GDP—that is, factoring in a population growth that in most countries was accelerating, because death rates had finally begun to fall sharply while birth rates remained high. But there were clear differences in economic performance among countries. Brazil, with a diversified economic base and much the largest internal market, and Panama, with its canal-based service economy, posted the best records, their GDP per capita doubling between 1950 and 1970; Mexico and Venezuela did almost as well, as did Costa Rica. But the Argentine economy seemed to stagnate, and few countries scored significant gains. Moreover, the conviction eventually grew in countries where ISI had been vigorously pushed that the easy gains in replacement of imports were coming to an end and that, to maintain adequate growth, it would be necessary to renew emphasis on exports as well. World market conditions were favourable for a revival of export promotion; indeed, international trade had begun a rapid expansion at the very time that inward-directed growth was gaining converts in Latin America.

Industrial exports

The promotion of industrial exports was slow to appear in Latin America. Brazil was the most successful, selling automobiles and automotive parts mainly to other developing countries but at times even to the industrialized world. A slightly less satisfactory alternative was the setting up of plants to assemble imported parts into consumer goods that were immediately exported, thus taking advantage of Latin America's low labour costs. Such plants proliferated along Mexico's northern border (where they were known as *maquiladoras*) but sprang up also in Central America and around the Caribbean. In other instances Latin Americans tried to develop new ("nontraditional") primary commodity exports. Colombian cut flowers were a highly successful example of the latter, promoted from the late 1960s through special incentives such as tax rebates; Colombia became the world's second leading flower exporter. It also assumed a leading role in the illicit narcotics trade. It enjoyed a brief boom of marijuana exports in the 1970s and in the following decade became the world's leading supplier of cocaine, which was processed in clandestine Colombian laboratories from coca leaf paste mostly originating in Bolivia and Peru.

Developments in social policy. Continued advances in public health were the principal basis for the explosion of population growth, which in turn made more difficult the provision of other social services. Nevertheless, educational coverage continued to expand, and state schools increased their share of students at the expense of private (often church-affiliated) institutions. Social security systems were introduced in countries that previously had none and expanded where they already existed. Yet such benefits chiefly went to organized urban workers and members of the middle sectors so that the net effect was often to increase, rather than lessen, social inequality.

Moreover, structural land reform received more lip service than actual implementation. Extensive land distribution did occur in Bolivia following that country's 1952 revolution, and in Cuba large private estates were eliminated after 1959; but Mexico, which had been the leader in this area, now tended to favour capitalist agribusinesses rather than peasant communities. The poor were also hurt by the high inflation that in the 1950s and after became endemic in Brazil and the Southern Cone and was intermittently a problem elsewhere, resulting in considerable part from an inability or unwillingness to generate by taxation the fiscal resources needed for economic and social development programs.

The United States and Latin America in the Cold War era. Whatever policies Latin-American countries adopted in the postwar era, they had to take into account the probable reaction of the United States, now more than ever the dominant power in the hemisphere. It was the principal trading partner and source of loans, grants, and private investment for almost all countries, and Latin-American leaders considered its favour worth having. Policy makers in Washington, on their part, were unenthusiastic about ISI and state-owned enterprises, but, as long as North American investors were not hindered in their own activities, the inward-directed policy orientation did not pose major problems. Moreover, as the Cold War developed between the United States and the Soviet Union, the great majority of Latin-American governments sided willingly with the former. Although they complained of being neglected by Washington's preoccupation with the threat of communism in Europe and Asia, a number of Latin-American countries offered at least token military collaboration during the Korean War.

A threat developed in Central America when the Guatemalan government of Jacobo Arbenz (1951–54), which frankly accepted the support of local communists, attacked the holdings of the United Fruit Company as part of an ambitious though ultimately abortive land reform. This combined political and economic challenge caused the United States to assist Guatemalan counterrevolutionaries and neighbouring Central American rulers in overthrowing Arbenz. The reversion to interventionist tactics featured use of the Central Intelligence Agency (CIA) rather than landing of military forces. But it foreshadowed later CIA assistance to the Chilean military in ousting their country's Marxist president, Salvador Allende, in 1973, not to mention the U.S. vendetta against the Sandinista revolutionary government that took power in Nicaragua in 1979, only to be worn down by covert action and economic harassment to the point that it peacefully accepted defeat in a free election in 1990.

Renewed U.S. interventionist policies

Impact of the Cuban Revolution. By most social and economic indicators, Cuba by mid-century was among Latin America's most highly developed countries. However, in the postwar period it was afflicted with lacklustre economic growth and a corrupt political dictatorship set up in 1952 by the same Batista who earlier had helped put his country on a seemingly democratic path. It was also a country whose long history of economic and other dependence on the United States had fed nationalist resentment, although control of the sugar industry and other economic sectors by U.S. interests was gradually declining. While conditions for revolutionary change were thus present, the particular direction that Cuba took owed much to the idiosyncratic genius of Fidel Castro, who, after ousting Batista at the beginning of 1959, proceeded by stages to turn the island into the hemisphere's first communist state, in close alliance with the Soviet Union.

The Cuban Revolution achieved major advances in health and education, though frankly sacrificing economic efficiency to social objectives. Expropriation of most private enterprise together with Castro's highly personalistic dictatorship drove many members of the middle and upper classes into exile, but a serious decline in productivity was offset for a time by Soviet subsidies. At the same time, thanks to its successful defiance of the United States—which tried and failed to overthrow it by backing a Cuban exiles' invasion in April 1961—and its evident social advances, Castro's Cuba was looked to as a model throughout Latin America, not only by established leftist parties but also by disaffected students and intellectuals of mainly middle-class origin.

Over the following years much of Latin America saw an upsurge of rural guerrilla conflict and urban terrorism, in response to the persistence of stark social inequality and political repression. But this upsurge drew additional inspiration from the Cuban example, and in many cases Cuba provided training and material support to guerrillas. The response of Latin-American establishments was twofold and eagerly supported by the United States. On one hand, governments strengthened their armed forces, with U.S. military aid preferentially geared to counterguerrilla operations. On the other hand, emphasis was placed on land reform and other measures designed to eliminate the root causes of insurgency, all generously aided by the United States through the Alliance for Progress launched by President John F. Kennedy.

Even though much of the reactive social reformism was cosmetic or superficial, the counterrevolutionary thrust was nonetheless generally successful. A Marxist, Salvador Allende, became president of Chile in 1970, but he did so by democratic election, not violent revolution, and he was overthrown three years later. The only country that appeared to be following the Cuban pattern was Nicaragua under the Sandinista revolutionary government, which in the end could not withstand the onslaughts of its domestic and foreign foes. Moreover, the Cuban Revolution ultimately lost much of its lustre even in the eyes of the Latin-American left, once the collapse of the Soviet Union caused Cuba to lose its chief foreign ally. Although the U.S. trade embargo imposed on Cuba had been a handicap all along, shortages of all kinds became acute only as Russian aid was cut back, clearly revealing the dysfunctional nature of Castro's economic management.

Political alternatives. *Movement toward democracy.* The Latin-American countries that did not opt for the Cuban model followed widely varying political paths. Mexico's unique system of limited democracy built around the Institutional Revolutionary Party was shaken by a wave of riots in the summer of 1968 on the eve of the Olympic Games held in Mexico City, but political stability was never seriously in doubt. A somewhat analogous regime was devised in Colombia as a means of restoring civilian constitutional rule after a brief relapse in the mid-1950s into military dictatorship: the dominant Liberal and Conservative parties chose to bury the hatchet, creating a bipartisan coalition (called the National Front) whereby they shared power equally between themselves while formally shutting out any minor parties. Once this arrangement expired in 1974, Colombia became again a more conventional political democracy, such as Costa Rica had been since before 1950 and Venezuela became in 1958 after the overthrow of its last military dictator.

Practice of democracy

In Latin America generally, the practice of democracy was somewhat sporadic, but, wherever regular elections took place, they involved an enlarged electorate. The last Latin-American countries adopted woman suffrage in the 1950s, and literacy tests continued to fall (as did illiteracy itself). Women also began to occupy high political office, including the presidency in Argentina (1974–76) and Bolivia (1979–80), though nowhere by popular election until Violeta Chamorro won the Nicaraguan vote of 1990 that put an end to Sandinista rule.

The advent of populism. The amorphous phenomenon of populism was another feature of the mid-20th-century political scene. Its consummate practitioner was Juan Perón of Argentina, who as member of a military regime that seized power in 1943 took a special interest in social policy. Perón wooed Argentine labour by means of wage increases and bonuses, pensions and fringe benefits, while also exploiting widespread resentment of an oligarchy that in the 1930s had reasserted its political as well as economic dominance. He promised social justice without violent class struggle and national greatness on the basis of industrial and military strength. His message, delivered in popular language, won Perón a clearcut victory when he ran for president in 1946.

Perón was not the first Latin-American leader to reward his followers with social benefits or to rail against native oligarchs and foreign imperialists, but he established a personal, charismatic link with ordinary citizens in a way that no one before him had done as successfully. Nor did he pretend to be heading a revolution. As president, aided by his wife Evita until her death in 1952, he continued to cultivate mass support but signally neglected to lay a sound basis for long-term economic growth. In any case numerous imitators and parallel figures appeared in postwar Latin America.

The leading party of postdictatorial Venezuela, Democratic Action (Acción Democrática; AD), was basically reformist in orientation but with populist overtones. Rómulo Betancourt and other AD leaders were less personalistic in style than Perón, who was finally overthown in 1955, but like him they stood for the granting of lavish benefits to the working and middle classes within a general framework of capitalism. In Venezuela oil wealth ultimately encouraged the national government to squander resources without adequate regard for the future. A similar charge was leveled against Juscelino Kubitschek, who became president of Brazil (1956–61) through his skill at old-style machine politics. He was technically not a populist but had the same bent for extravagant promises and freewheeling expenditure. Kubitschek's best-known achievement was the building of Brasília, the architecturally striking but fabulously expensive new capital city. Its construction aggravated inflationary woes but nicely epitomized his pledge to bring "fifty years' progress in five."

Brasília

Christian Democracy. A new feature since World War II was the appearance of a number of Christian Democratic parties, which offered a program of moderate reform inspired by Roman Catholic social teachings. Most were small splinter groups, but Christian Democrats eventually achieved power in Venezuela, El Salvador, and Chile. In Venezuela they alternated with the social democratic AD and in their policies became almost indistinguishable from it. In El Salvador in the 1980s they were enmeshed in a preexisting struggle against leftist guerrillas. In Chile, where they came to power first, under President Eduardo Frei (1964–70), they launched an ambitious land reform and partially nationalized the copper industry. They received enthusiastic support from the United States via the Alliance for Progress as presenting a promising alternative to Cuban-style revolution, but they failed to extend their mandate, going down to narrow defeat in a three-way contest won by Salvador Allende.

Bureaucratic authoritarianism. Allende as president combined Marxist assault on the owners of the means of production with populist lavishing of short-term benefits on his working-class followers, and on both counts he stirred violent resentment among upper- and middle-class Chileans as well as attracting the adamant hostility of the United States. In September 1973 he was ousted in favour of General Augusto Pinochet, who proved the most successful exponent of a new style of military dictatorship defined by political scientists as bureaucratic authoritarianism. It was not, of course, a complete novelty. It reflected the 20th-century Latin-America-wide phenomenon whereby the leadership of increasingly professionalized armies passed to sons of the middle class who had a commitment to modernizing the infrastructure of their societies. Such earlier dictatorships as that of Carlos Ibáñez del Campo (1927–31) during another Chilean relapse from constitutional rule had shown marked developmentalist tendencies. Bureaucratic authoritarianism, however, as practiced in Brazil after the coup of 1964, in Argentina by officers dedicated to keeping the Peronistas from regaining power, or in Chile under Pinochet, was a response to the perceived mismanagement of the economy by populists and other demagogues. It rested on the conviction that no democratically elected regime could afford to take the harsh measures needed to curb inflation, reassure foreign and domestic investors, and thereby quicken economic growth to the point that untrammeled democracy could be safely practiced. While military men kept order with varying degrees of harshness and human rights violations, civilian economists and technocrats would direct most other policy—whence the term bureaucratic authoritarianism.

Under Pinochet, the guiding voice in Chilean economic matters was assigned to a group of economists who had been trained at the University of Chicago and were strongly influenced by the monetarist school of Milton Friedman. Political authoritarianism stood in apparent contradiction to the generally free-market, laissez-faire policies prescribed in economic and social affairs; and, though inflation fell sharply, industrial production also dropped with the decline in the level of official protection. A similar combination of approaches arose under the military governments in Argentina in the 1960s and again from 1976 to 1983 and in Uruguay after 1973, again with mixed economic results. In Brazil from 1964 to 1985 military presidents and their technocratic advisers assigned a larger role in economic affairs to the state, while a Peruvian military regime that took power in 1968 undertook a radical program of social and economic reforms, giving way to a more typical bureaucratic-authoritarian regime only after running into serious economic difficulties. In all these countries, political repression fell lightly on most of the population, but anyone suspected of engaging in—or simply encouraging—active resistance was liable to arrest, torture, and in extreme cases forced "disappearance"; this was a notable feature of the last Argentine military regime. Moreover, military rule of one sort or another did spread until by 1980 democratically elected civilian governments could be found only in Colombia, Venezuela, Costa Rica, and (by stretching the definition just a bit) Mexico.

LATIN AMERICA AT THE END OF THE 20TH CENTURY

The last two decades of the 20th century witnessed a generalized economic crisis in Latin America, triggered in large part by external factors but aggravated by domestic mismanagement; in search of a way out, countries put their trust in neoliberal approaches favouring a free flow of trade and investment and reduction of the role of the state, all as recommended by the International Monetary Fund or other lending and advisory agencies. Even Castro's Cuba hesitantly embarked on the neoliberal economic path, though Castro did not show equal enthusiasm for the parallel political tendency, which was a turn to democratic procedures.

Neoliberal economic trends

Debt crisis. The ingredient of economic crisis that attracted widest attention was Latin America's inability to maintain full service on its foreign debt, which had grown to dangerously high levels. Both Mexico and Venezuela, as major petroleum exporters, benefited from rising international oil prices during the 1970s, but, instead of concluding that foreign credit was no longer necessary, they assumed that any amount of indebtedness would be easy to pay back. Brazil's generals drew a similar conclusion from their country's better-than-average economic growth. Even where no such circumstances were present, foreign private and institutional lenders had lost their depression-induced caution in lending to Latin America, and they had at their disposal an ever-greater flood of dollars to be placed in world financial markets. Bankers used often aggressive tactics in pressuring Latin-American governments to borrow, and the region's total foreign debt increased from 1970 to 1980 by more than 1,000 percent.

Developments in the world economy soon brought Latin America a rude awakening. Whereas commodity prices were generally favourable in the 1970s, a world recession in the following decade caused them to fall sharply. At the same time, interest rates rose in the United States and western Europe as governments sought to curb inflationary pressures and make other difficult adjustments. Latin America thus faced an increased debt bill, with fewer resources to pay it. Colombia alone managed to avoid default or compulsory rescheduling, and all countries faced severe fiscal problems. Domestic expenditures had to be cut back or financed through unsupported issues of paper money. Most of Latin America experienced slow or negative economic growth, together with inflation; indeed, hyperinflation was the rule in Argentina and Brazil and in some smaller countries. Real wages fell everywhere except Colombia and Chile.

Return to democracy. Latin America's democracies, and quasi-democratic Mexico, were politically less vulnerable to economic hard times than the dictatorships: their governments could be and were changed by regular electoral procedures, whereas dictatorial regimes that encountered similar problems had to be removed by other means. Armed force, however, was seldom necessary, and in Argentina change came from outside, in the form of Great Britain's embarrassing defeat of the Argentine military government's 1982 attempt to reoccupy the Falkland (Malvinas) Islands that Britain had seized a century and a half before. That fiasco completed the discrediting of the

Argentine regime and forced it to reinstate elective civilian government sooner than intended. Elsewhere, the force of domestic opinion, reinforced by foreign disapproval and by sheer discouragement on the part of ruling military officers, was usually enough to bring about a peaceful transition to democracy. By the beginning of the 1990s, elected civilian government was overwhelmingly the rule and dictatorship a rare exception.

The triumph of neoliberalism. One of the last countries to return to democracy was Chile, where the Pinochet dictatorship had been more successful than most in economic management. After first imposing harsh readjustments and committing its share of mistakes, it had launched the country on a steady course of economic growth that made it a much-admired model in Latin America and continued even after the dictator finally turned over the presidency (though not control of the armed forces) to an elected Christian Democrat in 1990. The Chilean model was based, in any event, on the application of neoliberal policies—reduction of trade barriers, privatization of state companies, encouragement of foreign as well as domestic private investment, and lessening of regulation generally—that to one degree or another were ultimately adopted by all countries, including (within limits) the surviving communist dictatorship of Cuba.

The Chilean model

A clear example of the new approach to economic affairs was Mexico's joining Canada and the United States in the North American Free Trade Agreement (NAFTA), which went into effect in 1994. Intra-Latin-American free trade arrangements moved forward too, and country after country sought private buyers for inefficient state firms. Several nations, led by Chile, moved to privatize in whole or in part their social security systems. There were nevertheless certain limits to neoliberal reforms: in Mexico and Venezuela, for example, state oil firms were exempt from the privatization process. Neither did bureaucracies and government expenditures shrink very rapidly, if they shrank at all.

Results of the new economic policies were, not surprisingly, mixed. Latin America remained vulnerable to the vagaries of world markets, and the greater opening to international trade often led to a dangerous upsurge in nonessential imports. The region's share of world exports hovered around 4 percent, less than half what it had been in 1950. As Latin America entered the final decade of the century, however, most countries saw a resumption of modest economic growth after the disastrous 1980s, an often dramatic decline in inflation, and a strengthening of the private sector of the economy. Socially, the changes produced both winners and losers, and in Mexico they aggravated the plight of Indian peasants in the southern state of Chiapas, who unleashed a renewal of guerrilla insurgency just as the country was entering NAFTA. Yet Latin America's revitalized commitment to political democracy—which did not mean sudden elimination of all human rights abuses and other deficiencies any more than in the rest of the world—appeared to face few serious challenges.

Religious trends. Roman Catholicism continued to be a powerful force in the second half of the 20th century. Its influence could be seen in the continuing prohibition, almost everywhere, of abortion and in the tendency to play down official support (which nevertheless existed) for birth control campaigns. Relations of the Roman Catholic church with the state and with society at large were meanwhile affected, however, by new currents within the church itself. The movement of renewal and reform undertaken by the second Vatican Council (1962–65) favoured mainstream Catholic teaching and practice at the expense of popular "folk Catholicism" yet led to a somewhat more tolerant approach toward other denominations. In addition, coinciding as it did with the impetus given to leftist movements by the Cuban Revolution, the call for renewal inspired an influential minority of priests and nuns to seek a synthesis of religious faith and political commitment under the banner of liberation theology. Some priests actually joined guerrilla bands, while others laboured to "raise the consciousness" of their flocks concerning social injustice. This brand of activism met with general disapproval from Latin-American governments, especially military regimes, some of which brutally persecuted the clergy involved. It also divided the church, and without gaining the widespread popular allegiance that "liberationist" clergy had hoped for.

Liberation theology

In the late 20th century the principal religious development was a rapid expansion of Protestantism, especially the Evangelical and Pentecostal churches. With a primary emphasis on individual spiritual improvement and salvation and a closeness between ministers and laity that neither traditional nor renewed Catholicism could match, the Protestants rapidly increased their numbers throughout Latin America. In countries as diverse as Brazil and Guatemala there were by the end of the century more Protestants than actively churchgoing Roman Catholics. Protestantism was not strong among traditional elites or in intellectual circles, but its adherents were beginning to attain positions of influence.

A changing society. Despite the expansion (sometimes impressive, sometimes not) of the middle strata of Latin-American society, by the late 20th century progress toward reducing historically high levels of social inequality was disappointing almost everywhere save in communist Cuba. The poorest countries of western Europe enjoyed greater per capita income than the wealthiest in Latin America. Yet, with regard to such social indicators as literacy and life expectancy, Costa Rica, Cuba, and the nations of the Southern Cone approximated the standards of the industrialized world, and for Latin America as a whole the lag was substantially less than in 1900 or 1950.

The rate of population growth, having peaked in the third quarter of the century, fell to about 2 percent a year, though with wide variations among countries. In parts of northern Latin America a factor contributing to this decline was emigration to the more prosperous and politically stable United States, whose major metropolitan areas, Los Angeles and Miami above all, held large and growing Latin-American communities. Total population was still approaching a half billion, of whom a substantial majority were urban dwellers; Latin America contained two of the world's largest cities, Mexico City and São Paulo, each approaching 20 million inhabitants. However, primate cities grew more slowly than intermediate centres. In the cities, where literacy and then access to television were nearly universal, people were exposed more and more quickly to new trends and ideas emanating from the United States or western Europe; to a lesser degree the same forces, and the continuing improvement of road transportation, were also decreasing the isolation of rural Latin Americans.

With social and economic modernization came changes, too, in gender relations. In most of Latin America women achieved full legal equality with men only gradually and usually later than winning the vote. In Argentina, for example, wives gained equal authority with husbands over minor-aged children only after the return to democracy in the 1980s. However, women did take advantage of increased educational and employment opportunities to gain more control of their own lives. Almost as many women as men were enrolled in secondary education, and the traditional alternatives of domestic service and prostitution for those who chose or were obliged to work outside the home had been supplemented by an array of clerical, professional, and light factory jobs. Falling birth rates gave further evidence that women were pursuing new options. The fact that domestic servants were still relatively inexpensive made it easier for middle- and upper-class women to pursue professional careers. Servants, however, were less inclined than they once were to accept their position as permanent; realistically or not, they dreamed of something better and to that extent epitomized a more general yearning for personal and social improvement that posed a challenge for all Latin-American nations. (D.Bu.)

Gender relations

BIBLIOGRAPHY

General works. LESLIE BETHELL (ed.), *The Cambridge History of Latin America* (1984–), is a general reference work. EDWIN WILLIAMSON, *The Penguin History of Latin America* (1992); and SIMON COLLIER, HAROLD BLAKEMORE, and THOMAS E. SKIDMORE (eds.), *The Cambridge Encyclopedia of Latin America and the Caribbean*, 2nd ed. (1992), offer introductory material. TULIO HALPERÍN DONGHI, *The Contemporary History of Latin America* (1993; originally published in Spanish, 1970), focuses on the region's colonial and neocolonial relations with North Atlantic nations.

Early Latin America. An overview is presented in JAMES LOCKHART and STUART B. SCHWARTZ, *Early Latin America: A History of Colonial Spanish America and Brazil* (1983), emphasizing analysis over narrative detail and treating all of Latin America as a unit.

Spanish America: C.H. HARING, *The Spanish Empire in America* (1947, reprinted 1985), is an institutionalist classic. SHERBURNE F. COOK and WOODROW BORAH, *Essays in Population History*, 3 vol. (1971–79), is by the field's most illustrious demographers.

WILLIAM HICKLING PRESCOTT, *History of the Conquest of Mexico*, 3 vol. (1843), and *History of the Conquest of Peru*, 2 vol. (1847), both available in many later printings, are outdated but still read. LEWIS HANKE, *The Spanish Struggle for Justice in the Conquest of America* (1949, reprinted 1965), concentrates on the ideological crusades of fray Bartolomé de las Casas. JAMES LOCKHART, *Spanish Peru, 1532–1560: A Social History*, 2nd ed. (1994), surveys Hispanic conquest society in a central area. IDA ALTMAN, *Emigrants and Society: Extremadura and America in the Sixteenth Century* (1989), follows people across the Atlantic. ROBERT RICARD, *The Spiritual Conquest of Mexico* (1966; originally published in French, 1933), is a masterpiece of institutionalist historiography but ignores the roles of the indigenous people and the Spanish civil population.

Studies of indigenous society include CHARLES GIBSON, *The Aztecs Under Spanish Rule* (1964), a large work based mainly on Spanish sources; WILLIAM B. TAYLOR, *Drinking, Homicide & Rebellion in Colonial Mexican Villages* (1979); NANCY M. FARRISS, *Maya Society Under Colonial Rule* (1984, reissued with corrections, 1992), a broad treatment combining historical and anthropological techniques; JAMES LOCKHART, *The Nahuas After the Conquest* (1992), based largely on sources in Nahuatl; and KAREN SPALDING, *Huarochirí: An Andean Society Under Inca and Spanish Rule* (1984). FREDERICK P. BOWSER, *The African Slave in Colonial Peru, 1524–1650* (1974), thoroughly investigates the role of Africans in a Spanish American area. P.J. BAKEWELL, *Silver Mining and Society in Colonial Mexico: Zacatecas, 1546–1700* (1971), broadly treats a major silver mining centre. MURDO J. MACLEOD, *Spanish Central America: A Socioeconomic History, 1520–1720* (1973, reprinted 1984), is the cornerstone of early Guatemalan historiography. ROBERT J. FERRY, *The Colonial Elite of Early Caracas: Formation & Crisis, 1567–1767* (1989), investigates society in a fringe area. D.A. BRADING, *Miners and Merchants in Bourbon Mexico, 1763–1810* (1971), is a massive social and economic study of Mexico's late-colonial international economy and government, and *Haciendas and Ranchos in the Mexican Bajío: León, 1700–1860* (1978), shows the rationality and market orientation of the agricultural sector. JOHN K. CHANCE, *Race and Class in Colonial Oaxaca* (1978), studies urban demography. JOHN E. KICZA, *Colonial Entrepreneurs: Families and Business in Bourbon Mexico City* (1983–), broadly surveys urban society. ERIC VAN YOUNG, *Hacienda and Market in Eighteenth-Century Mexico: The Rural Economy of the Gudalajara Region, 1675–1820* (1981), links the growth of agrarian estates to the size and nature of urban populations. SUSAN MIGDEN SOCOLOW, *The Merchants of Buenos Aires, 1778–1810: Family and Commerce* (1978), is a prosopographical study.

Brazil: ALEXANDER MARCHANT, *From Barter to Slavery: The Economic Relations of Portuguese and Indians in the Settlement of Brazil, 1500–1580* (1942, reissued 1966), though badly outdated in its ethnohistorical aspect, shows typical European economic procedures in a fringe area. GILBERTO FREYRE, *The Masters and the Slaves (Casa-Grande & Senzala): A Study in the Development of Brazilian Civilization*, 2nd ed., rev. (1946, reissued 1986; originally published in Portuguese, 4th ed., 2 vol., 1943), concerns the society of northeastern Brazil in the time of sugar production. A.J.R. RUSSELL-WOOD, *Fidalgos and Philanthropists: The Santa Casa da Misericórdia of Bahia, 1550–1755* (1968), is tantamount to a study of northeastern urban society and partly updates Freyre. STUART B. SCHWARTZ, *Sovereignty and Society in Colonial Brazil: The High Court of Bahia and Its Judges, 1609–1751* (1973), combines social and institutional approaches, and his *Sugar Plantations in the Formation of Brazilian Society: Bahia, 1550–1835* (1985), embraces many methods, materials, and topics. C.R. BOXER, *The Golden Age of Brazil, 1695–1750* (1962, reissued 1995), extracts a maximum of social, economic, and general information from institutional-narrative materials. (J.Lo.)

Independence to 1910. *General works:* DAVID BUSHNELL and NEILL MACAULEY, *The Emergence of Latin America in the Nineteenth Century*, 2nd ed. (1994), focuses particularly on politics. E. BRADFORD BURNS, *The Poverty of Progress: Latin America in the Nineteenth Century* (1980), creatively argues that modernization hurt the majority of Latin Americans.

Spanish America: JOHN TUTINO, *From Insurrection to Revolution in Mexico: Social Bases of Agrarian Violence, 1750–1940* (1986), reviews rural rebels' motives from the wars preceding independence to those of the Mexican Revolution. JOHN LYNCH, *The Spanish American Revolutions, 1808–1826*, 2nd ed. (1986), is the best account of the political and military events of the wars for independence. VERENA MARTINEZ-ALIER (VERENA STOLCKE), *Marriage, Class, and Colour in Nineteenth-Century Cuba*, 2nd ed. (1989), links race, gender, and class factors. REBECCA J. SCOTT, *Slave Emancipation in Cuba* (1985), portrays the abolition of slavery in Cuba as the result of social, political, and economic factors. PAUL GOOTENBERG, *Between Silver and Guano: Commercial Policy and the State in Postindependence Peru* (1989), examines the complicated social contests through which integration into global economic relations emerged after independence. JONATHAN C. BROWN, *A Socioeconomic History of Argentina, 1776–1860* (1979), challenges assumptions about Latin America's economic dependence on North Atlantic powers. FLORENCIA E. MALLON, *The Defense of Community in Peru's Central Highlands: Peasant Struggle and Capitalist Transition, 1860–1940* (1983), details the social and political aspects of a period of economic overhaul. CHARLES A. HALE, *Mexican Liberalism in the Age of Mora, 1821–1853* (1968), provides a model of intellectual history.

Brazil: EMILIA VIOTTI DA COSTA, *The Brazilian Empire* (1985; originally published in Portuguese, 1977), collects well-written essays on liberalism, slavery, the end of the empire, and other major topics. WARREN DEAN, *Rio Claro: A Brazilian Plantation System, 1820–1920* (1976), examines changing labour arrangements during the growth and demise of slavery. JOÃO JOSÉ REIS, *Slave Rebellion in Brazil: The Muslim Uprising of 1835 in Bahia* (1993; originally published in Portuguese, 1986), analyzes one of the century's most influential revolts. STANLEY J. STEIN, *Vassouras: A Brazilian Coffee County, 1850–1900* (1985), studies the rise and decline of a traditional coffee zone. LINDA LEWIN, *Politics and Parentela in Paraíba: A Case Study of Family-Based Oligarchy in Brazil* (1987), analyzes in detail the patronage system that defined politics in late 19th- and early 20th-century Brazil. THOMAS H. HOLLOWAY, *Immigrants on the Land: Coffee and Society in São Paulo, 1886–1934* (1980), analyzes how two million European immigrants worked and lived. (R.A.Ki.)

Twentieth-century Latin America. *General works:* THOMAS E. SKIDMORE and PETER H. SMITH, *Modern Latin America*, 4th ed. (1997); and PETER CALVERT and SUSAN CALVERT, *Latin America in the Twentieth Century*, 2nd ed. (1993), are good introductions to the period.

Political developments: Regional treatments of political topics include A.E. VAN NIEKERK, *Populism and Political Development in Latin America* (1974; originally published in Dutch, 1972); and JOHN A. PEELER, *Latin American Democracies: Colombia, Costa Rica, Venezuela* (1985), assessing variants of liberal democracy. The impact of the Cuban Revolution and associated leftist currents is explored in JORGE G. CASTAÑEDA, *Utopia Unarmed: The Latin American Left After the Cold War* (1993); and WILLIAM E. RATLIFF, *Castroism and Communism in Latin America, 1959–1976* (1976). FREDERICK M. NUNN, *The Time of the Generals: Latin American Professional Militarism in World Perspective* (1992), discusses the political role of the military.

Economics: VICTOR BULMER-THOMAS, *The Economic History of Latin America Since Independence* (1994), provides the best overall treatment. Labour issues are examined in HOBART A. SPALDING, JR., *Organized Labor in Latin America* (1977), a helpful introduction; and CHARLES BERGQUIST, *Labor in Latin America* (1986). ALAIN DE JANVRY, *The Agrarian Question and Reformism in Latin America* (1981), covers the history of agrarian reform.

International relations: BILL ALBERT and PAUL HENDERSON, *South America and the First World War* (1988); and R.A. HUMPHREYS, *Latin America and the Second World War*, 2 vol. (1981–82), examine the repercussions of foreign conflicts. Good examples from the vast literature examining the role of the United States are BRYCE WOOD, *The Dismantling of the Good Neighbor Policy* (1985); LESTER D. LANGLEY, *The United States and the Caribbean in the Twentieth Century*, 4th ed. (1989); BARBARA STALLINGS, *Banker to the Third World: U.S. Portfolio Investment in Latin America, 1900–1986* (1987); and FREDRICK B. PIKE, *FDR's Good Neighbor Policy: Sixty Years of Generally Gentle Chaos* (1995).

Culture and society: Useful studies include JEAN FRANCO, *The Modern Culture of Latin America*, rev. ed. (1970); FRANCESCA MILLER, *Latin American Women and the Search for Social Justice* (1991); and DAVID STOLL, *Is Latin America Turning Protestant? The Politics of Evangelical Growth* (1990). (D.Bu.)

Latin-American Literature

Latin-American literature encompasses the national literatures of the Spanish-speaking countries of the Western Hemisphere and Portuguese-speaking Brazil. Historically, it also includes the literary expression of the highly developed American Indian civilizations conquered by the Spaniards. Over the years, Latin-American literature has developed a rich and complex diversity of themes, forms, creative idioms, and styles. A concise survey of its development is provided here.

The article is divided into the following sections:

THE COLONIAL YEARS (1492–1826)

Literature of the conquest (1492–1600). With the discovery of new lands beyond the ocean, Spain and Portugal embarked on a venture that created empires and stamped the colonial seal on vast areas of the Americas. This adventure was chronicled from the day Christopher Columbus set sail; his letters to King Ferdinand and Queen Isabella of Spain marked the beginnings of a rich body of colonial writings. The discovery and conquest are told in countless letters, chronicles, histories, polemical tracts, dictionaries, grammars, religious pieces, and epic poems. This corpus was the voice of the colonizers and represented European intellectual influences, shaped by medieval and Renaissance doctrines. The cultures discovered and conquered by the Spaniards also possessed a rich heritage of poetry, theatre, and mythico-historical writing, the most poignant of which are their chronicles of the conquest and of their defeat and destruction. Recent research has brought forward the voice of the defeated, or those previously marginalized from standard discourse, such as the Native American chronicler Felipe Guamán Poma de Ayala and the Native American historians.

The excitement of the first contact and the adventures and problems that followed are recorded in various writings, notably the five dispatches, or *Cartas de relación de Fernando Cortés sobre el descubrimiento y conquista de la Nueva España* (1519–26; "Letters of Relation of Fernando Cortés on the Discovery and Conquest of New Spain"; Eng. trans. *Five Letters, 1519–1526*), sent by Hernán Cortés to his emperor, Charles V; the indignant account of Bernal Díaz del Castillo, who spoke for the common soldier in *Historia verdadera de la conquista de la Nueva España* (1632; "The True History of the Conquest of New Spain"); the impassioned pages of the *Brevísima relación de la destruyción de las Indias* (1552; "Very Brief Account of the Destruction of the Indians"; Eng. trans. *The Tears of the Indians*) by the Dominican friar Bartolomé de las Casas, which became the cornerstone of the "black legend" of colonial Spain; and the observations on the tragic results of the struggle between cultures expressed in the *Comentarios reales, que tratan del origen de los Incas* (1609–17; "Royal Commentaries that Treat of the Origin of the Incas") by Garcilaso de la Vega, a mestizo (person of European and Indian parentage) born of an Incan princess and a Spanish conqueror. Vega was trained in humanist doctrines and eager to recoup the knowledge of the lost empire.

Following their Renaissance-European models and fueled by the heroic fervour of their age and by the example of Ariosto, Tasso, and the Latin masters, the colonizers created verse epics narrating their adventures in the American vastness. *La Araucana* (1569–89; *The Araucaniad*), the first and the best of these epics, was begun on scraps of paper and bark by the young captain and courtier Alonso de Ercilla y Zúñiga while he was serving with the Spanish armies fighting the Araucanians of Chile. Its great attraction is still the immediacy of the experience and the unexpectedly balanced attitude toward the Indians. Enormously popular, *La Araucana* inspired similar efforts elsewhere: Juan de Castellanos, *Elegías de varones ilustres de Indias* (1589; "Elegies of the Illustrious Gentlemen of the Indies"); Martín del Barco Centenera, *La Argentina, o la conquista del Rio de la Plata* (1601; *The Argentine and the Conquest of the River Plate*); and Gaspar Pérez de Villagrá, *Historia de la Nueva México* (1610; *History of New Mexico*). These later poems pale in poetic import but are of documentary interest. Only the *Arauco domado* (1596; *Arauco Tamed*) of Pedro de Oña, Chile's first outstanding native poet, approaches Ercilla's epic in artistic achievement. Epic beginnings

Bernardo de Balbuena (Mexico) best represents the rich and varied literary expression of the dawning Baroque age. His epic of Spanish history, *El Bernardo, o victoria de Roncesvalles* (1600; "Bernard, or Victory of the Roncesvalles"), and his eclogues are overshadowed by his *Grandeza mexicana* (1604; "The Grandeur of Mexico"), a paean to the attractions of the new metropolis of Mexico City.

The discovery and conquest of the tropical heartland of America produced no rich body of Brazilian counterparts to the spirited epics in prose and verse of the Spanish conquerors. The Portuguese found no centuries-old native civilizations upon which to build a new society. Indians were few, and, rather than resist, they retreated into the interior, leaving the invaders to settle at widely scattered points along the extended narrow coastal plain. The beautiful and bountiful coast and forest lands set the tone even today for *brasilidade,* the love of Brazil.

The Brazilian equivalent of Cortés' letters is found in the letter of discovery sent to King Manuel I of Portugal by the scribe Pero Vaz de Caminha, who accompanied the explorer Pedro Álvares Cabral in 1500. The missionary labours were largely in the hands of the Jesuits, and one of their number, José de Anchieta, is considered the father of Brazilian literature. The epic *Prosopopéa* (1601; "Prosopopoeia"), by Bento Teixeira Pinto, is commonly regarded as the first book by a Brazilian-born writer. Origin of Brazilian literature

The flowering of colonial letters (1600–1808). By 1600 the day of the conqueror was past, and the New World entered a quieter phase as colonial society stabilized. The first printing press in the Americas was established in Mexico City about 1539, another in Lima in 1584. Both cities were granted university charters in 1551. In Brazil no university was founded and no printing permitted until the close of the colonial period. Everywhere in the Americas a growing leisure class devoted more time to intellectual and artistic pursuits, and close ties with the homeland encouraged the development of parallel literary patterns. Life became a reflection of trends across the Atlantic, although the majority of the populace had little or no economic stability. Frequent outbursts of racial and religious persecution produced a literature of the oppressed that is still being discovered, and the imprint of Indian traditions can be found among even cultivated writers.

The theatre, emerging from medieval forms at the time of the conquest, served during the 16th century as a missionary medium in the Indian tongues for the conversion of the Native Americans. There were also a few secular representations in the Indian languages of preconquest memories (some of these works were written by Creole authors—*i.e.,* authors with European background), and in some remote areas preconquest traditions continued to be staged well into the 19th century. However, for the bulk of the non-Indian population—consisting of Creoles and urbanized mestizos—the dramatic repertoire was mainly Spanish and Portuguese. Plays written in America were patterned after the drama of the Spanish Golden Age, and the greatest of the native dramatists, the Mexican Juan Ruiz de Alarcón, even lived and wrote in Spain.

Fiction was banned by the Spanish crown—although clandestine imported works circulated widely—and emerged only with independence, leaving the field in prose to historical and biographical works. One of the best was *El lazarillo de ciegos caminantes* (*c.* 1773; "A Guide for Inexperienced Travelers"), a satire in the form of a travelogue of a journey from Buenos Aires to Lima. Its author, Alonso Carrió de la Vandera, was a Spanish postal official who used the pseudonym Concolorcorvo to protect himself from official reprisal. In a culture where the distinction between history and literature was all but nonexistent, many of the chronicles and historical works contained early fictional elements.

Poetry was the most popular form of literature, and a contest held in Mexico in 1585 attracted more than 300 mostly indifferent representatives of the Spanish fashion. One late-17th-century poet distinguished herself above all in the colonial world: the Mexican nun Sor Juana Inés de la Cruz. Her metaphysical poem *Primero sueño* (1692; "First Dream") and, above all, her beautiful, profane lyrics, *Poemas de la unica poetisa americana, musa dezima, Soror Juana Inez de la Cruz* (1690; "Poems of the Only American Poetess, Tenth Muse, Sister Juana Inés de la Cruz"; Eng. trans. *Poems*), are among the greatest literary achievements of Latin America. She also wrote a considerable body of scientific speculation, now lost, and a number of plays; and her response to criticisms by the bishop of Puebla, *Respuesta a Sor Filotea de la Cruz* (1691; "Reply to Sister Filotea of the Cross"; Eng. trans. *A Woman of Genius*), is an extraordinary intellectual document.

During the colonial period there appeared a satiric tradition protesting against social and individual suffering in such authors as the Peruvian Juan del Valle y Caviedes and Gregório de Matos Guerra, the epigrammatic "Devil's mouthpiece" of Brazil, both influenced by the Spanish satirist Francisco Gómez de Quevedo y Villegas. Their verse was widely circulated and avidly read. Toward the end of the colonial period, this style developed popular forms that stirred the revolutionary spirit.

Literature of rebellion (1808–26). Political unrest spread in Latin America as the 18th century advanced, and French ideas broke through the controls held by Spain and Portugal over New World thought. Early stirrings became more purposeful under the impact of the French and American revolutions. Printing presses and periodicals sprang up, and ideals quickened in literary societies founded by Latin America's young liberals.

Emergence of literary societies

One of the earliest fruits of contact with foreign thought was the abortive Brazilian Inconfidência Mineria of 1789, a "conspiracy of poets" headed by Joaquim José da Silva Xavier and supported by a number of exceptional writers who were members of the Minas school of epic and Neoclassical poets. This school had no equal in the Spanish colonies of that day. José Basílio da Gama's *O Uraguai* (1769; *Uruguay*) and José de Santa Rita Durão's *Caramuru: poema épico do descobrimento da Bahia* (1781; "Epic Poem of the Discovery of Bahia") were two native epics of Brazil. Love of country and the appearance of the Indian as a literary character marked both works as forerunners of Brazilian intellectual independence as well as of incipient Romanticism. When, after a bloodless victory in 1822, Brazil emerged as an empire, José Bonifácio de Andrade e Silva stood out as a patriarch of Brazil's struggle for independence. Author of vigorous, passionate verse (*Poesias,* 1825), he is considered Brazil's first Romantic poet.

Francisco de Miranda, a Venezuelan leader of Spanish-American independence, left a remarkable journal that revealed the influence of his contacts in the United States. His compatriot Simón Bolívar was later christened "the thinker of the Revolution" for his prophetic analyses of the sociopolitical scene. Bolívar's writings exhibit some of the best of French thought, particularly that of Charles-Louis de Secondat, Baron de Montesquieu, and of Jean-Jacques Rousseau. But the Mexican José Joaquín Fernández de Lizardi, referred to as *el pensador mexicano* ("the Mexican thinker"), is the only liberator remembered as a man of letters, primarily for a picaresque tale, *El periquillo sarniento* (1816; *The Itching Parrot*), considered the first Latin-American novel.

The revolution found popular expression in balladry and heroic verse. More enduring patriotic poetry came with victory. José Joaquín Olmedo, an Ecuadorean, limited his work almost exclusively to the themes and spirit of the revolutionary years. His best-known poem, *La victoria de Junín, canto a Bolívar* (1826; "The Victory of Junín, Song to Bolívar"), is a fine example of heroic poetry in the Classical style. In his Virgilian *Silva a la agricultura de la zona tórrida* (1826; "Silva to the Agriculture of the Torrid Zone"; Eng. trans. *A Georgic of the Tropics*), Andrés Bello of Venezuela exhorted his fellow Americans to turn their swords into plowshares and to cultivate the natural riches of America. In later years Bello distinguished himself as a grammarian, critic, lawgiver, and educator and as the intellectual father of Chile's Romantic generation. José María Heredia of Cuba lived most of his life in political exile (Venezuela, New York City, Mexico), where the grandeur of nature and the Indian past inspired his widely acclaimed poems *En el teocalli de Cholula* (1820; "On the Teocalli of Cholula") and *Oda al Niágara* (1824; "Ode to Niagara"). Although still largely Neoclassical in form, his poetry foreshadowed the new literary movement of Romanticism with which the young Latin-American nations were born into political independence.

Patriotic poetry

THE 19TH CENTURY

Romanticism. Political independence from Spain and Portugal did not bring an end to despotism, anarchy, and repression of the Indian and African populations; economic and political stability for most new nations came late and with difficulty, if at all. In literature, American themes fired the imagination of the liberators, but Neoclassical forms were still dominant. European Romanticism also pointed the way to cultural independence, even though that way lay largely along a route marked out by French, English, and Spanish writers. The controlled form of Neoclassicism yielded to the freedom, individualism, and emotional intensity of Romanticism, and the European cult of the medieval became in many cases a passion for the Indian—present and past. The most illustrious early Romantics were Argentine political refugees who fled from the dictator Juan Manuel de Rosas. Their leader was Esteban Echeverría, who—after a stay in France (1826–30), where Romanticism was at its height—led the movement in *La cautiva* (1837; "The Captive"), one of the earliest fusions of native themes and scenes with newer free verse forms. The polemical Domingo Faustino Sarmiento, like so many of Latin America's 19th-century intellectuals a statesman, educator, and writer, provided the first serious study of the *pampas,* or plains, and of gaucho (cowboy) lore in his *Facundo* (1845; *Life in the Argentine Republic in the Days of the Tyrants, or Civilization and Barbarism*), a biographical mélange of fiction, sociology, and politics written in passionate denunciation of Rosas.

This emphasis on the national scene gave birth to an indigenous literary genre without European prototype, the gaucho literature of Argentina and Uruguay. The gaucho had long been the subject of folk literature and soon figured in cultivated verse, as in Rafael Obligado's poem *Santos Vega* (1887), on the legendary minstrel, and the humorous *Fausto* (1866; *Faust*) by Estanislao del Campo, until finally he received epic treatment in *Martín Fierro* (1872–79; *The Gaucho Martin Fierro*) by José Hernández.

Emergence of gaucho literature

This Romantic evocation of national themes and types reached its poetic climax while also pointing the way to the next period in the elegiac *Tabaré* (1886; *Tabaré: An Indian Legend of Uruguay*) by the Uruguayan Juan Zorrilla de San Martín. Zorrilla's epic poem related the fate of the aboriginal Charrúas, who were vanquished by Spanish invaders. In prose, Juan León Mera of Ecuador contributed *Cumandá, o, un drama entre salvages* (1871; "Cumandá, or, A Drama Among Savages"), and Manuel de Jesús Galván of the Dominican Republic added *Enriquillo: leyenda histórica dominicana* (1879–82; "Enriquillo: Dominican Historical Legend"; Eng. trans. *The Cross and the Sword*) to a growing number of fictionalized portrayals of Indian life. The Colombian Jorge Isaacs wrote the popular, lachrymose idyll *María* (1867), and, among the growing number of political novels, José Mármol gave a one-sided view of life in Argentina under the tyranny of Rosas in *Amalia* (1851–55). The *tradición,* patterned after the historical anecdote, was exploited by the Peruvian Ricardo Palma, whose series called *tradiciones peruanas* ("Peruvian traditions") appeared between 1872 and 1910.

The Brazilian Romantics had no similar past to idealize and instead extolled the natural beauties of their homeland and the simple Indian life. Domingo José Gonçalves de Magalhães, though still eclectic, launched the Romantic movement with *Suspiros poéticos e saudades* (1836; "Poetic Sighs and Longings"). The best and most representative of Brazil's Romantic poets is Antônio Gonçalves Dias, whose best-known poem, "Canção do Exílio" (1843; "Song of Exile"), laments his exile from his country and family. Later phases were exemplified in a poetry of doubt and despair by Manuel Antônio Alvares de Azevedo, author of *Noite na taverna* (1855; "Night at the Tavern"), and in the sociopolitical verse of Antônio de Castro Alves, author of *Os escravos* (1876; "The Slaves").

The theatre was not one of Latin America's strongest genres during the Romantic period, consisting primarily of melodrama, regional comedy, and political pamphleteering. Among the exceptions are the Neoclassical comedies of Manuel Eduardo de Gorostiza (Mexico), the satires of Manuel Ascensio Segura (Peru), and the dramas Cuba's Gertrudis Gómez de Avellaneda created out of serious psychological analysis and careful dramatic structure. She produced her plays in Spain, however, where she spent most of her life.

Realism and naturalism. Romantic interest in the picturesque and unusual led writers to discover a regional way of life, and shortly after midcentury the *cuadro de costumbres,* or sketch of contemporary customs, developed into a realistic novel of manners, often with an urban setting. Many of these novels are the expression of an ideal national history connected to political independence. From that time, the novel assumed a more commanding role in Latin-American letters, but it appeared almost concurrently with the several types representative of successive literary trends in Europe, whose masters provided the molds into which Latin-American themes were poured. Alberto Blest Gana began producing a series of costumbrista novels on Chilean life, of which *Martín Rivas* (1862) was considered the best.

Cuadro de costumbres

Naturalism in the manner of Émile Zola made its appearance with the Argentine Eugenio Cambaceres. After midcentury, several late-Romantic political writers, both Brazilian and Spanish-American, distinguished themselves in essay form. Among them were Juan Montalvo, Eugenio María de Hostos, Joaquim Nabuco, and Rui Barbosa. Manuel González Prada, an ironic experimental poet and essayist, was the chief figure, and his verse paved the way for the new poetry of a coming generation of rebels.

The true novel appeared first in Brazil in 1844 with *A moreninha* ("The Little Brunette") by Joaquim Manuel de Macedo. Still one of the most widely read of his country's novelists, José Martiniano de Alencar initiated a vogue of the Brazilian Indianist novel with *O Guaraní* (1857; "The Guaraní Indian") and *Iracema* (1865). These romantic tales of love between Indian and white, however, represented only one aspect of Alencar's varied literary activity. He also turned to the life and customs of Brazil's backlands, and *O gaúcho* (1870) and *O sertanejo* (1875; "The Man of the Backlands"), though still markedly Romantic in spirit, were among the forerunners of a flourishing regional literature. Two other contributors to this transitional genre were Alfredo d'Escragnolle Taunay, whose *Inocência* (1872) became a universal favourite, and Bernardo Guimarães, whose abolitionist *A escrava Isaura* (1875; "The Slave Isaura") was a decisive step toward the novel of social protest.

True realism, with a definite leaning toward naturalism, began with *Memórias de um sargento de milícias* (1854; "Memories of a Sergeant of the Militias") by Manuel Antônio de Almeida, but it was not until the mid-1870s that the novel began to expose widespread social and psychological maladjustments to a rapidly changing economic scene. Aluízio Azevedo, an early example of social protest in the manner of 20th-century novelists, wrote such favourites as *O mulato* (1881; "The Mulatto") and *O cortiço* (1890; *A Brazilian Tenement*). Less occupied with external aspects of Brazilian life, Joaquim Maria Machado de Assís pried into the psychological complex of the Brazilian and distinguished himself as his country's most original and gifted writer. His trilogy, *Memórias póstumas de brás cubas* (1881; *Epitaph of a Small Winner*), *Quincas Borba* (1891; *Philosopher or Dog?*), and *Dom Casmurro* (1899; Eng. trans. *Dom Casmurra*), was a landmark in Latin-American letters.

True realism

Rubén Darío and the Modernists. In Spanish America a measure of political and economic stability aided a cosmopolitan rebellion against the sentimental romanticism that filled the literary pages of the newspapers and magazines. Young writers across the Americas immersed themselves in the mainstream of world thought and writing. Somewhat disparagingly labeled Modernists (Modernistas) by the older generation, they wrote on exotic themes, often shutting themselves off from their immediate environment in artificial worlds of their own making—the ancient past, the distant Orient, and the lands of childhood fancy and sheer creation. Beauty was their goddess and "art for art's sake" their creed. Influenced by French movements, they followed no regular path; Symbolism, Parnassianism, Decadentism, and all the rest coexisted in any given individual or followed each other in any order.

Foremost among the early Modernists were the Mexican Manuel Gutiérrez Nájera, whose elegiac verse and agile rhythmic prose sketches and tales best represented the transition from the excesses of Romanticism to the more filigreed "Modernism" (Modernismo); the Colombian José Asunción Silva, who wrote a small but influential body of savagely ironic and elegiac poems; the Cuban Julián del Casal, cultivator of the exquisite Parnassian sonnet; and his compatriot José Martí, martyr and symbol of Cuba's struggle for freedom from Spain, whose inspired prose style and deceptively simple, sincere verse set his work above and apart from all schools and movements.

The full flowering of Modernism, however, came under the leadership of one of the greatest poets in Spanish, Rubén Darío of Nicaragua. His collection of verse and prose, *Azul* ("Blue"), published in 1888, pointed the initial way, but *Prosas profanas* (1896; "Lay Hymns") represented the high point of the cosmopolitan phase of the movement. Darío blended Modernist formal experimentation with an expression of inner despair or an almost metaphysical joy in *Cantos de vida y esperanza* (1900; "Songs of Life and Hope"). When Spain's empire crumbled in 1898 and mutual sympathy allayed the old distrust between Spain and its former colonies, he turned to Hispanic traditions as he had so often turned to Hispanic forms. In the face of U.S. imperialism, he spoke for Hispanic solidarity. Darío's imitators, particularly of his early experimental, escapist phase, were often slavish copiers, but Darío and his fellow Modernists brought about the greatest revitalization of language and poetic technique in Spanish since the 17th century. Many of his contemporaries were writers of considerable merit: Mexico's Amado Nervo, whose Orientally influenced mysticism was reflected in *Serenidad* (1914; "Serenity") and *Elevación* (1900; "Elevation"); Peru's José Santos Chocano, whose exalted Americanism gave birth to *Alma América* (1906; "American Soul"); Bolivia's Ricardo Jaimes Freyre, who

Literary revival

drew upon Scandinavian mythology for *Castalia bárbara* (1899; "The Barbarous Castalia"); Colombia's Guillermo Valencia, whose classic bent was manifest in *Ritos* (1898; "Rituals"); and Uruguay's philosopher and essayist José Enrique Rodó, whose *Ariel* (1900) distinguished him as the leading theoretician and exponent of Modernist ideals.

Even after Darío's death in 1916, the majority of Spanish-American Modernists continued to be spellbound by the verbal magic and brilliance of his *Prosas profanas*. There were, however, other prominent figures: Leopoldo Lugones of Argentina underwent drastic shifts ranging from Laforguian irony through Baroque complexity to an intense nationalism expressing itself in poems based on popular folk songs; Enrique González Martínez of Mexico wrestled with social and ethical problems in Symbolist sonnets; and Julio Herrera y Reissig of Uruguay was perhaps the outstanding Symbolist of Modernism, finding new visions in a landscape represented in startling images.

In Brazil, reaction against Romantic verse never produced the rich mosaic of Spanish-American Modernism. The Brazilian Parnassians, with their formalistic, detached poetry, were challenged by poets attracted to French Symbolism, but a Symbolist movement as such never materialized. Parnassianism, epitomized in the poetry of Raimundo Correia, Alberto de Oliveira, and Olavo Bilac, was unchanged until the developments of the 1920s altered Brazilian letters drastically.

LITERARY DEVELOPMENTS OF THE 20TH CENTURY

World Wars I and II, the intervening worldwide economic depression of the 1930s, and the Spanish Civil War thrust the nations of Latin America onto the international scene, altering the perspective and general orientation of Latin-American men and women of letters. Regionalistic preoccupations gave way to more universal concerns, a shift that was accompanied not only by new themes but also by new literary modes and stylistic techniques. The works that emerged during the second half of the century demonstrate the full maturing of Latin-American literature and its entry into the mainstream of Western letters.

Vanguard literature. When the Modernist leaders Darío and Lugones turned to frequent use of more traditional forms and especially to themes dealing with the troubled external world, their followers and many younger poets continued to explore the complex versification and language and new thematic sources with which Modernism had enriched Hispanic poetry. Prominent among them was an extraordinary group of women, some of whom dealt with traditional themes. Love in its various manifestations, maternal longing, and social protest were the subject matter of the Uruguayans Delmira Agustini and Juana de Ibarbourou and of the Chilean Gabriela Mistral, winner of the 1945 Nobel Prize for Literature. The Argentine Alfonsina Storni dealt in an innovative, ironic idiom with personal anguish and the difficulties of being a woman in a stern male world. Springing from various aspects of the Modernist tradition were the allusive, interiorized lyrics of the Peruvian José María Eguren; the idiosyncratic blend of daring imagery, Baudelairean struggle, and provincial themes of the Mexican Ramón López Velarde; and the brilliant sonnet sequence of the tropical scene in *Tierra de promisión* (1921; "Land of Promise") by the Colombian José Eustasio Rivera.

Poetry after World War I was extraordinarily rich and complex. Some groups, following the nihilistic waves of post-World War I "isms," experimented with free verse, often daring to use obscure imagery that gave a mistaken impression of a coldly intellectual mood. Many of these vanguard experimenters wrote out of a sociopolitical commitment as well. In the experiments of the Puerto Rican Luis Palés Matos, the Cuban Nicolás Guillén, and others, the voice and song of the African tradition were carried to a high artistic level.

Vicente Huidobro of Chile is important for his insistence on the poet's total creation of an autonomous world, an approach that was extremely influential on younger figures. Argentina's Jorge Luis Borges launched Ultraísmo in Buenos Aires in 1921 but evolved from these avant-garde beginnings to a poetry reflecting love for his city, a familiarity with numerous foreign literatures, and a preoccupation with metaphysical themes. These latter elements are also discernible in Borges' short stories, which brought him international fame and have influenced writers throughout the world. César Vallejo of Peru fused social concerns with Surrealism and the heritage of Modernism to create an intensely subjective and often obscure but vital poetry. The Chilean Pablo Neruda, winner of the 1971 Nobel Prize for Literature, also blended Marxism with Surrealism in his earlier work; he attempted a poetic synthesis of the suffering of the Americas in the *Canto general* (1950; "General Song"). In his enormous production Neruda made poetry of even the smallest aspects of the world he saw about him. There were numerous other distinguished poets of the Postmodernist period; the years from the beginning of Modernism to the end of Postmodernism rival the Spanish Golden Age of the 16th and 17th centuries in wealth of poetry.

Borges and Ultraísmo

The vanguard revolt in Brazil, usually referred to as Modernism (Modernismo; although quite different from Spanish-American Modernism), broke away noisily from academicism and colonial cultural bondage at the noted Modern Art Week program in São Paulo in 1922. The primary aim of the Brazilian Modernists (Modernistas) was to modernize national thought and life, casting aside the persistent vestiges of the 19th century. This high-pitched, often theatrical, self-searching period of aesthetic reevaluation and analysis of the immediate Brazilian present served as a necessary purge and produced such important figures in Brazilian literature as the movement's leader, Mário de Andrade, a gifted poet and musicologist; his lieutenant, Oswald de Andrade; Ronald de Carvalho, a critic and poet; and Manuel Bandeira, who has been acclaimed the country's greatest modern lyric poet. Preoccupation with social and metaphysical problems and an imperative urge toward untrammeled self-expression characterized the poetry of Modernist contemporaries or followers, such as Jorge de Lima, Cecília Meireles, and Augusto Federico Schmidt.

Voyage to the self. After Modernism, many Spanish Americans returned to the difficult realities of their nations for their subjects. The Modernists Martí and González Prada were hailed as intellectual progenitors, men aware of their responsibilities in guiding the Americas in a rapidly changing world. Rufino Blanco Fombona of Venezuela assailed his country's tyrants in *El hombre de hierro* (1907; "The Man of Iron") and *El hombre de oro* (1916; *The Man of Gold*). He was also among the first to attack U.S. imperialism, abetted by a militant Argentine, Manuel Ugarte. The latter's *El porvenir de América latina* (1911; "The Future of Latin America") and *El destino de un continente* (1923; *The Destiny of a Continent*), together with the writings of the Peruvian Francisco García Calderón, envisioned Latin America as the future guardian of the Latin tradition. Ricardo Rojas, an Argentine literary historian and critic, and José Vasconcelos, a controversial Mexican philosopher and educator, were more concerned with racial and cultural aspects within the American family of nations. Alfonso Reyes, a Mexican poet, scholar, and critic, made the essay a vital, intimate force and raised it to a new level of artistic excellence.

It was in prose fiction that the American scene in all its drama was best described. The disheartening years following the abolition of slavery in 1888 and the establishment of a republic in 1889 made serious-minded Brazilians analyze their troubled homeland as an extraordinary amalgam of people, land, and climate. Euclides da Cunha revealed the bedrock of Brazilian life in the epic *Os sertões* (1902; "The Backlands"; Eng. trans. *Rebellion in the Backlands*); this work was the first written protest on behalf of Brazil's forgotten man, the emerging Brazilian of the backlands. In *Canaã* (1902; *Canaan*), a novel of ideas, José Pereira da Graça Aranha focused on the effects of recent European immigration upon this evolving Brazilian type. José Bento Monteiro Lobato's collection of short stories *Urupês* (1918; "Shelf Fungi") showed that intellectuals were still probing for native traits in a search that gave direction to the Brazilian Modernist impetus of the 1920s.

A new cultural regionalism of Brazil's "Northeastern school" that flowered after 1930 produced gifted prose writers, including a sociologist, Gilberto Freyre, whose *Casa-grande & senzala* (1933; "Big House and Slave Barrack"; Eng. trans. *The Masters and the Slaves*) was fundamental to an understanding of the region. José Lins do Rego, in a highly personal, evocative style, depicted the clash of the old and new way of life in his classic "Sugarcane" cycle and in *Pedra bonita* (1938; "Beautiful Stone"); Jorge Amado gave Brazil some of Latin America's best proletarian literature in *Cacau* (1933; "Cacao"), *Jubiabá* (1935), and *Terras do sem fim* (1942; *The Violent Land*). *Angústia* (1936; *Anguish*), by Graciliano Ramos, reflected the impact of European literature's fascination with the individual inner struggle. Érico Veríssimo was one of Latin America's most distinguished cosmopolitan writers, as demonstrated in *Olhai os lírios do campo* (1938; *Consider the Lilies of the Field*) and *O tempo e o vento* (1950; *Time and the Wind*).

New trends in the Spanish-American novel

Contrary to its exceptional development in Brazil, the novel in Spanish America had been left largely to a few Modernist novelists, such as Manuel Díaz Rodríguez of Venezuela and Enrique Larreta of Argentina, and to the followers of the French naturalist Émile Zola. Baldomero Lillo of Chile gave artistic dimension to the sufferings of oppressed miners in *Sub sole* (1907; "Under the Ground"); and Federico Gamboa of Mexico, in *Santa* (1903), dealt with a prostitute reminiscent of Zola's protagonist in *Nana* from a Roman Catholic point of view—the only version of naturalism possible in the Hispanic world. Later novelists, essayists, and short-story writers developed newer and more effective techniques under the influence of various foreign innovators, including Marcel Proust, James Joyce, Franz Kafka, and William Faulkner.

The horror and bloodshed of the Mexican Revolution (1910–17) shocked much of the complacent intellectual minority into a realization of the plight of their country's submerged masses: the Indian and mestizo found champions. The most widely acclaimed work resulting from the Mexican Revolution was the novel *Los de abajo* (1915; *The Underdogs*) by Mariano Azuela, an army doctor of one of the bands of the revolutionary general Pancho Villa. Azuela chronicled the revolution from the point of view of the humble peasants who were its soldiers and its victims. From the late 1920s on, Martín Luis Guzmán, Gregorio López y Fuentes, and José Rubén Romero, among many others, gave further breadth and scope to this turbulent period but never equaled Azuela's treatment.

The Indian, as represented in 20th-century literature, was not the uncompromising hero of the epic *La Araucana* nor the symbol of colonial revolt against tyrannical Spain and much less the "noble savage" of the untamed romantic wilderness. He was rather the victim of political and economic forces that kept the masses in abject bondage to colonial institutions. This new *indianista* literature had its roots in such novels as *Aves sin nido* (1889; *Birds Without a Nest*) by Clorinda Matto de Turner of Peru. The Indian's cause, moreover, was advanced by González Prada, a precursor of the militant pro-Indian, social reform party APRA (American Popular Revolutionary Alliance) in 1923. The most extreme example of this committed literature was the brutally realistic school of Ecuadorean writers. Its most influential member, Jorge Icaza, produced mass-directed, vernacular novels such as *Huasipungo* (1934; *Huasipungo: The Villagers*) and *En las calles* (1935; "In the Streets").

Major themes

The clash between the forces of nature and powerful economic pressures, and between the land and the city, had not died with the passing of 19th-century local chiefs and remained an important literary theme. The land and its resources were more zealously sought than ever before, and the struggle became more violent as man fought against man for the possession of these riches. In the growing urban centres, modern industrial economies exerted an even more insidious control over the destiny of the masses. The clash between the old and the new on the pampas of Argentina and Uruguay inspired the sombre descriptive pages of Uruguay's short-story writer Javier de Viana; the psychological portrayal of rural types in *El terruño* (1916; "The Native Country") by Carlos Reyles, also of Uruguay; the brilliant imagery of the mythic re-creation of the gaucho, *Don Segundo Sombra* (1926; *Don Segundo Sombra: Shadows on the Pampas*) by Ricardo Güiraldes of Argentina; and the tragic depths of the irreconcilable difference between the city and pampa in *El inglés de los güesos* (1924; "The Englishman of the Bones") by another Argentine, Benito Lynch. In *Doña Bárbara* (1929) Rómulo Gallegos gave a dramatic depiction of similar forces at work on the Venezuelan llanos, or plains.

Two artists were representative of a growing number of younger writers who had discovered the selva, or tropical rain forest: Horacio Quiroga of Uruguay, a consummate short-story artist who excelled both in fantasy and in dramatic descriptions of the struggle for life in the jungle of Misiones, Arg.; and José Eustasio Rivera, a Colombian poet whose sole prose work, *La vorágine* (1924; *The Vortex*), was a powerful denunciation of exploitation in the inner Amazon basin during the rubber boom of the early 1900s.

In *La maestra normal* (1900; "The Grade-School Teacher") Manuel Gálvez, an Argentine novelist, captured the pettiness and monotony of smaller provincial centres before modern mechanized manners shattered old colonial ways. Chile's poor had two champions in the novelist Joaquín Edwards Bello and the short-story writer Manuel Rojas. Man's struggle with the deeper forces within and beyond himself was re-created in the psychological novels of the Chilean Eduardo Barrios and in the tales of the Cuban Alfonso Hernández Catá.

Recent trends. The literature of the last half of the 20th century has been characterized by an increased preoccupation with man as the victim of alienating forces—solitude, identity crises, anguish, and evil—and by a marked determination to create new forms and techniques. Above all, it has put forth a new language more responsive to the demands imposed by increasingly complex spiritual, social, and ideological concerns. These concerns found a direct expression in the essay, a form that has been cultivated with distinction by writers equally competent in other genres. Individual and collective preoccupation with analysis ranged from the ruthless dissection of his country's ills by the Argentine Ezequiel Martínez Estrada in *Radiografía de la pampa* (1933; *X-Ray of the Pampa*) to the more social and psychological probings of the Mexican Octavio Paz (*El laberinto de la soledad* [1950; *The Labyrinth of Solitude*]). Paz, recipient of the Nobel Prize for Literature in 1990, is also notable as a perceptive analyst of world poetry and as a poet who wedded the heritage of Surrealism with Indian and popular traditions and with Oriental mysticism. Many poets veered from experimental and hermetic emphasis to seek answers to the same concerns in a more personalized style, often striving for a direct, conversational, and sometimes heavily ironic tone. Nicanor Parra (Chile) popularized a similar style in several volumes, notably *Poemas y antipoemas* (1954; *Poems and Antipoems*). Younger figures such as José Emilio Pacheco (Mexico) relied on a conversational tone and explored concepts of poetry as ever-changing dialogue with the reader and as a product remote from any hypothetical author.

Experimentation in the theatre

The drama had enjoyed considerable popularity in the larger cities since the late 1800s, although relatively few plays were of real note. By the end of the century, Cuba and Argentina both had a flourishing satiric popular theatre based on regional types and language. In Argentina and Uruguay this led, at the turn of the century, to an important regional realistic theatre anchored in social problems, particularly the complex impact of intense immigration. Its outstanding practitioner was the Uruguayan Florencio Sánchez. The late 1920s saw an outburst of experimental activity, reflecting influences ranging from Expressionism to Eugene O'Neill, Luigi Pirandello, and Jean Cocteau. Since that time, the Latin-American theatre has absorbed forces as varied as Arthur Miller, Tennessee Williams, Samuel Beckett, and Eugène Ionesco. Among the best of those leading the movement of renovation were the Brazilian Joracy Camargo (*Deus lhe pague* [1932; "God Pays Them"]) and the Mexican Rodolfo Usigli (*El gesticulador* [1937; *The Gesticulator*] and *Corona de sombra: pieza an-*

tihistórica en tres actos [1943; *Crown of Shadows: An Anti-Historical Play*]), a caustic commentator of his compatriots' foibles and a seeker of a national self-comprehension. A younger generation—including Sebastián Salazar Bondy of Peru, Emilio Carballido of Mexico, Osvaldo Dragún and Griselda Gambaro of Argentina, Jorge Díaz and Egon Wolff of Chile, José Triana of Cuba, and Ariano Suassuna of Brazil—won international recognition for its technical expertise and its humane responses to social problems.

After the unrest of the late 1960s, the Latin-American theatre became increasingly politicized, and many groups under the inspiration of Enrique Buenaventura of Colombia and Augusto Boal of Brazil developed a theory and practice of theatre as a collective creation, although more recently this collective emphasis has lessened considerably. More recent dramatists have continued experimenting widely with popular and nontraditional forms while retaining an emphasis on the complex social realities of Latin America: Chileans like Carlos Cerda, whose *Lo que está en el aire* (1986; "What Is in the Air") employed avant-garde techniques to portray the terror of the dictatorship; the Argentines who explored the difficulties of the years of the dictators and the return to a freer but perhaps equally unstable society (*e.g.*, Eduardo Pavlovsky, *El señor Galíndez* [1986; "Mr. Galíndez"]); or the many Mexicans rethinking the nature of Mexican history, such as Vicente Leñero in *El juicio* (1972; "The trial").

Predominance of prose fiction

Prose fiction has occupied the centre of the contemporary Latin-American literary scene. This new narrative has given full voice to social concerns through techniques that launched an attack on both the form of the novel and the structure of the language. The new prose was already evident in the work of a small band of earlier writers, exemplified especially in the extraordinary tales of the Argentine Jorge Luis Borges (*El Aleph* [1949; *The Aleph and Other Stories, 1933–1969*, or *Labyrinths*]); in the terrifying portrait of tyranny *El señor presidente* (1948; *The President*), by the Guatemalan Miguel Angel Asturias, winner of the 1967 Nobel Prize for Literature; in the Joycean examination of the social and psychological roots of the revolution of 1910 by the Mexican Agustín Yáñez (*Al filo del agua* [1947; *The Edge of the Storm*]); and in the ideological and philosophical novels of the Argentine Eduardo Mallea (*La bahía de silencio* [1940; *The Bay of Silence*] and *Todo verdor perecerá* [1941; *All Green Shall Perish*]).

Later examples of this remarkable group include the vision of America as a new and different reality expressed in Baroque language by the Cuban Alejo Carpentier (*Los pasos perdidos* [1953; *The Lost Steps*]); the earthy, enigmatic prose of the Brazilian João Guimarães Rosa (*Grande sertão: veredas* [1956; "Big Backland: Paths"; Eng. trans. *The Devil to Pay in the Backlands*]); the Rabelaisian romances of a Brazilian known for his social commitment, Jorge Amado (*Gabriela, cravo e canela* [1958; *Gabriela, Clove and Cinnamon*]); the excursions into paranoia and apocalypse of *El túnel* (1948; *The Tunnel*) by Ernesto Sabato (Argentina); and the complex metaphors of salvation and damnation into which Juan Rulfo of Mexico transformed the arid rural landscape (*El llano en llamas* [1953; *The Burning Plain and Other Stories*] and *Pedro Páramo* [1955]).

Principal among the creators of a new Latin-American literature was the Argentine Julio Cortázar, who explored interlocking realities in works such as *Rayuela* (1963; *Hopscotch*), which challenged the whole concept of the structure of fiction. It was on the bedrock of such work that the younger generation planted the invigorating and challenging contemporary novel. Among the leaders of this generation was the Colombian Gabriel García Márquez, winner of the 1982 Nobel Prize for Literature for his mythic vision of Spanish-American society in novels such as *Cien años de soledad* (1967; *One Hundred Years of Solitude*). This vision was continued in his later work, including *El amor en los tiempos del cólera* (1996; *Love in the Time of Cholera*). Carlos Fuentes of Mexico examined the interaction between the present and the still-living past in volumes ranging from *Gringo viejo* (1985; *The Old Gringo*) to *La campaña* (1990; *The Campaign*), a "historical" novel re-creating the wars of independence of the early 19th century. The Peruvian Mario Vargas Llosa examined the social and moral forms of Spanish America in works such as *La ciudad y los perros* (1963; "The City and the Dogs"; Eng. trans. *The Time of the Hero*) and the later *La fiesta del chivo* (2000; "Feast of the Goat"), dealing with Rafael Trujillo's dictatorship in the Dominican Republic. The Cubans Guillermo Cabrera Infante and Reinaldo Arenas and the Puerto Rican Luis Rafael Sánchez, each in his own way, exposed the authoritarian systems pressuring the individual and the changing nature of Caribbean societies in novels of extreme linguistic daring, revolutionizing the structures of the language and of the novel to create highly personal visions. Other leaders were Chile's José Donoso, who examined the decline of feudal society in novels that range from realistic to bizarre, and Manuel Puig of Argentina, who studied the psychological impact of popular culture and psychosexual realities. His *El beso de la mujer araña* (1976; *The Kiss of the Spider Woman*) was internationally famous as a novel and later successful as a film and play.

Contemporary characteristics

The subsequent generation of writers reexamined accepted history and continued this concern with technique and language in novels of muralistic scope, vital exuberance, creative use of anachronism, and considerable linguistic and structural complexity. Eugenio Aguirre (Mexico) reassessed the contact between Spaniards and Maya in *Gonzalo Guerrero: novela histórica* (1980); in *Los perros del paraíso* (1983; *The Dogs of Paradise*), Abel Posse (Argentina) saw Columbus' voyage as analogous to Argentine history. Gustavo Sáinz (Mexico) in *Fantasmas aztecas* (1982; "Aztec Ghosts") examined coexisting ancient and modern national realities. There was a considerable social and political dimension to this new fiction, reflecting the uncertainties of the Cold War and post-Cold War periods as well as the ambiguous relations with the United States. Many authors struggled to represent the battle against local tyrants. At the same time, these novelists retained in the main a driving interest in technique, as in Ariel Dorfman (Chile), whose *Moros en la costa* (1973; *Hard Rain*) is a vision of political reality expressed in a mixture of different kinds of prose covering a wide range of writing styles and ultimately questioning the nature of writing itself. The same is true of the Brazilian writers Silviano Santiago (*Em liberdade* [1981; "In Liberty"]) and Rubem Fonseca (*Bufo & Spallanzani* [1985]).

The most significant development as Latin America approached the 21st century was the emergence of a host of recognized women writers. Among them were Chilean Isabel Allende, whose *La casa de los espíritus* (1982; *The House of Spirits*) was widely acclaimed; Mexican Laura Equivel (*Como agua para chocolate* [1989; *Like Water for Chocolate;* internationally successful as a film]); and Argentine Luisa Valenzuela (*Novela negra con argentinos* [1990; *Black Novel with Argentines*]). Following the example of writers such as Alfonsina Storni, a number of younger figures—among them the Mexican Angeles Mastretta (*Arráncame la vida* [1985; *Mexican Bolero*]), the Uruguayan Cristina Peri Rossi (*La última noche de Dostoievski* [1992; *Dostoevsky's Night*]), and the Puerto Ricans Ana Lydia Vega (*Falsas crónicas del sur* [1991; "False Chronicles of the South"]) and Rosario Ferré (*Flight of the Swan* [2001; written in English])—combined the double tendency toward metafiction and the denunciation of oppression with a strong feminist tone.

As shown by the importance of women writers, a notable aspect of the literature at the turn of the century was the emergence of new voices. They expressed, among other things, gender issues, the immigrant experience, and popular life and culture. These expressions were often postmodern in that they offered no one answer, either as to content or as to style. Cuban writer Severo Sarduy was an early, brilliant example of this trend, in such works as the experimental *De donde son los cantantes* (1967; *From Cuba with a Song*), which includes three narratives encompassing the entire history of Cuba, and the even more experimental *Cobra* (1972; Eng. trans., *Cobra*), where the setting is a transvestite theatre and some episodes occur in India and China.

(J.E.E./F.N.D./Ed.)

BIBLIOGRAPHY

General histories and reference works. A translation of the standard, reliable history, with somewhat dated value judgments, is ENRIQUE ANDERSON-IMBERT, *Spanish-American Literature: A History*, 2nd ed., rev., updated, and trans. by ELAINE MALLEY, 2 vol. (1969; originally published in Spanish, 6th ed., 1967). ROBERTO GONZÁLEZ ECHEVARRÍA and ENRIQUE PUPO-WALKER (eds.), *The Cambridge History of Latin American Literature*, 3 vol. (1996), covers the sweep of Latin-American literature from pre-Columbian times to late 20th century, including Brazil. WILLIAM LUIS (ed.), *Modern Latin-American Fiction Writers: First Series* (1992), vol. 113 of the *Dictionary of Literary Biography*, has life-and-works essays on major and minor authors. CARLOS A. SOLÉ and MARIA ISABEL ABREU (eds.), *Latin American Writers*, 3 vol. (1989), has generally good essays on authors from colonial times to late 20th century.

Anthologies. ROBERTO GONZÁLEZ ECHEVARRÍA (ed.), *The Oxford Book of Latin American Short Stories*, trans. from Spanish (1997, reissued 1999), has stories from pre-Columbian times to late 20th century and includes Brazil. EMIR RODRÍGUEZ MONEGAL and THOMAS COLCHIE (eds.), *The Borzoi Anthology of Latin American Literature*, 2 vol. (1977), has particularly good coverage of the 20th century. STEPHEN TAPSCOTT (ed.), *Twentieth-Century Latin American Poetry: A Bilingual Anthology* (1996), is a superb collection covering the late 19th to late 20th centuries, including Brazilian Portuguese poetry; it has an excellent introduction.

Colonial. ROBERTO GONZÁLEZ ECHEVARRÍA, *Celestina's Brood: Continuities of the Baroque in Spanish and Latin American Literatures* (1993), has chapters on the colonial Baroque. An overview of writing by women in the colonial period appears in JULIE GREER JOHNSON, *Women in Colonial Spanish American Literature: Literary Images* (1983).

Eighteenth century. A good primary source is JOSÉ JOAQUÍN FERNÁNDEZ DE LIZARDI, *The Itching Parrot*, trans. by EUGENE PRESSLY and ed. by KATHERINE ANNE PORTER (1942; originally published in Spanish, 1816). Wide-ranging if dated treatments of the Spanish-American Enlightenment are ARTHUR P. WHITAKER (ed.), *Latin America and the Enlightenment*, 2nd ed. (1961, reissued 1969); and A. OWEN ALDRIDGE (ed.), *The Ibero-American Enlightenment* (1971).

Romanticism to Modernism. ROBERTO GONZÁLEZ ECHEVARRÍA, *Myth and Archive: A Theory of Latin American Narrative* (1990, reissued 1998), has a chapter on travel literature and 19th-century narrative, particularly Sarmiento's. CATHY LOGIN JRADE, *Modernism, Modernity, and the Development of Spanish American Literature* (1998), is an excellent study of Modernism and its legacy. WILLIAM LUIS, *Literary Bondage: Slavery in Cuban Narrative* (1990), has important chapters on 19th-century antislavery narratives.

Poetry from the vanguard to the turn of the century. A volume that is good on the transition from Modernism to the vanguard is GWEN KIRKPATRICK, *The Dissonant Legacy of Modernismo: Lugones, Herrera y Reissig, and the Voices of Modern Spanish American Poetry* (1989). VICKY UNRUH, *Latin American Vanguards: The Art of Contentious Encounters* (1994), is comprehensive and authoritative on the various avant-garde movements and their manifestos.

Modern novel. The best study of regionalist fiction is CARLOS J. ALONSO, *The Spanish American Regional Novel: Modernity and Autochthony* (1990). SALVADOR BACARISSE (ed.), *Contemporary Latin American Fiction: Carpentier, Sabato, Onetti, Roa, Donoso, Fuentes, García Márquez* (1980), has some good essays on major figures. A good reference work is JOHN S. BRUSHWOOD, *The Spanish American Novel: A Twentieth-Century Survey* (1975). WILLIAM LUIS (ed.), *Voices from Under: Black Narrative in Latin America and the Caribbean* (1984), is a good discussion of blacks in literature and of literature by blacks. RAYMOND LESLIE WILLIAMS, *The Postmodern Novel in Latin America: Politics, Culture, and the Crisis of Truth* (1995, reissued 1997), has discussion by regions (Andean, Southern Cone, Caribbean), including late 20th-century novels. (Ed.)

Latin Literature

Latin literature was the product of the Roman Republic and the Roman Empire. When Rome fell, Latin remained the literary language of the Western medieval world until it was superseded by the Romance languages it had generated and by other modern languages. After the Renaissance the writing of Latin was increasingly confined to the narrow limits of certain ecclesiastical and academic publications. This article focuses primarily on ancient Latin literature. It does, however, provide a broad overview of the literary works produced in Latin by European writers during the Middle Ages and Renaissance.

The article is divided into the following sections:

Ancient Latin literature

Literature in Latin began as translation from the Greek, a fact that conditioned its development. Latin authors used earlier writers as sources of stock themes and motifs, at their best using their relationship to tradition to produce a new species of originality. They were more distinguished as verbal artists than as thinkers; the finest of them have a superb command of concrete detail and vivid illustration. Their noblest ideal was *humanitas*, a blend of culture and kindliness, approximating the quality of being "civilized."

Little need be said of the preliterary period. Hellenistic influence came from the south, Etrusco-Hellenic from the north. Improvised farce, with stock characters in masks, may have been a native invention from the Campania region (the countryside of modern Naples). The historian Livy traced quasi-dramatic *satura* (medley) to the Etruscans. The statesman-writer Cato and the scholar Varro said that in former times the praises of heroes were sung after feasts, sometimes to the accompaniment of the flute, which was perhaps an Etruscan custom. If they existed, these *carmina convivalia*, or festal songs, would be behind some of the legends that came down to Livy. There were also the rude verses improvised at harvest festivals and weddings and liturgical formulas, whose scanty remains show alliteration and assonance. The nearest approach to literature must have been in public and private records and in recorded speeches.

STYLISTIC PERIODS

Ancient Latin literature may be divided into four periods: early writers, to 70 BC; Golden Age, 70 BC–AD 18; Silver Age, AD 18–133; and later writers.

Early writers. The ground for Roman literature was prepared by an influx from the early 3rd century BC onward of Greek slaves, some of whom were put to tutoring young Roman nobles. Among them was Livius Andronicus, who was later freed and who is considered to be the first Latin writer. In 240 BC, to celebrate Rome's victory over Carthage, he composed a genuine drama adapted from the Greek. His success established a tradition of performing such plays alongside the cruder native entertainments. He also made a translation of the *Odyssey.* For his plays Livius adapted the Greek metres to suit the Latin tongue; but for his *Odyssey* he retained a traditional Italian measure, as did Gnaeus Naevius for his epic on the First Punic War against Carthage. Scholars are uncertain as to how much this metre depended on quantity or stress. A half-Greek Calabrian called Ennius adopted and Latinized the Greek hexameter for his epic *Annales,* thus further acquainting Rome with the Hellenistic world. Unfortunately his work survives only in fragments.

Greek influence of Livius Andronicus

The Greek character thus imposed on literature made it more a preserve of the educated elite. In Rome, coteries emerged such as that formed around the Roman consul and general Scipio Aemilianus. This circle included the statesman-orator Gaius Laelius, the Greek Stoic philosopher Panaetius, the Greek historian Polybius, the satirist Lucilius, and an African-born slave of genius, the comic playwright Terence. Soon after Rome absorbed Greece as a Roman province, Greek became a second language to educated Romans. Early in the 1st century BC, however, Latin declamation established itself, and, borrowing from Greek, it attained polish and artistry.

Plautus, the leading poet of comedy, is one of the chief sources for colloquial Latin. Ennius sought to heighten epic and tragic diction, and from his time onward, with a few exceptions, literary language became ever more divorced from that of the people, until the 2nd century AD.

Golden Age, 70 BC–AD 18. The Golden Age of Latin literature spanned the last years of the republic and the virtual establishment of the Roman Empire under the reign of Augustus (27 BC–AD 14). The first part of this period, from 70 to 42 BC, is justly called the Ciceronian. It produced writers of distinction, most of them also men of action, among whom Julius Caesar stands out. The most prolific was Varro, "most learned of the Romans," but it was Cicero, a statesman, orator, poet, critic, and philosopher, who developed the Latin language to express abstract and complicated thought with clarity. Subsequently, prose style was either a reaction against, or a return to, Cicero's. As a poet, although uninspired, he was technically skillful. He edited the *De rerum natura* of the philosophical poet Lucretius. Like Lucretius, he admired Ennius and the old Roman poetry and, though apparently interested in Hellenistic work, spoke ironically of its extreme champions, the *neōteroi* ("newer poets").

The Ciceronian period

After the destruction of Carthage and Corinth in 146 BC, prosperity and external security had allowed the cultivation of a literature of self-expression and entertainment. In this climate flourished the *neōteroi,* largely non-Roman Italians from the north, who introduced the mentality of "art for art's sake." None is known at first hand except Catullus, who was from Verona. These poets reacted against the grandiose—the Ennian tradition of "gravity"—and their complicated allusive poetry consciously emulated the Callimacheans of 3rd-century Alexandria. The Neoteric influence persisted into the next generation through Cornelius Gallus to Virgil.

Virgil, born near Mantua and schooled at Cremona and Milan, chose Theocritus as his first model. The self-consciously beautiful cadences of the *Eclogues* depict shepherds living in a landscape half real, half fantastic; these allusive poems hover between the actual and the artificial. They are shot through with topical allusions, and in the fourth he already appears as a national prophet. Virgil was drawn into the circle being formed by Maecenas, Augustus' chief minister. In 38 BC he and Varius introduced the young poet Horace to Maecenas; and by the final victory of Augustus in 30 BC, the circle was consolidated.

With the reign of Augustus began the second phase of the Golden Age, known as the Augustan Age. It gave encouragement to the classical notion that a writer should not so much try to say new things as to say old things better. The rhetorical figures of thought and speech were mastered until they became instinctive. Alliteration and onomatopoeia (accommodation of sound and rhythm to sense), previously overdone by the Ennians and therefore eschewed by the *neōteroi,* were now used effectively with due discretion. Perfection of form characterizes the odes of Horace; elegy, too, became more polished.

The Augustan Age

The decade of the first impetus of Augustanism, 29–19 BC, saw the publication of Virgil's *Georgics* and the composition of the whole *Aeneid* by his death in 19 BC; Horace's *Odes,* books I–III, and *Epistles,* book I; in elegy, books I–III of Propertius (also of Maecenas' circle) and books I–II of Tibullus, with others from the circle of Marcus Valerius Messalla Corvinus, and doubtless the first recitations by a still younger member of his circle, Ovid. About 28 or 27 BC Livy began his monumental history.

Maecenas' circle was not a propaganda bureau; his talent for tactful pressure guided his poets toward praise of Augustus and the regime without excessively cramping their freedom. Propertius, when admitted to the circle, was simply a youth with an anti-Caesarian background who had gained favour with passionate love elegies. He and Horace quarreled, and after Virgil's death the group broke up. Would-be poets now abounded, such as Horace's protégés, who occur in the *Epistles;* Ovid's friends, whom he remembers wistfully in exile; and Manilius, whom no one mentions at all. Poems were recited in literary circles and in public, hence the importance attached to euphony, smoothness, and artistic structure. They thus became known piecemeal and might be improved by friendly suggestions. When finally they were assembled in books, great care was taken over arrangement, which was artistic or significant (but not chronological).

Meanwhile, in prose the Ciceronian climax had been followed by a reaction led by Sallust. In 43 BC he began to publish a series of historical works in a terse, epigrammatic style studded with archaisms and avoiding the copiousness of Cicero. Later, eloquence, deprived of political influence, migrated from the forum to the schools, where cleverness and point counted rather than rolling periods. Thus developed the epigrammatic style of the younger Seneca and, ultimately, of Tacitus. Spreading to verse, it conditioned the witty couplets of Ovid, the tragedies of Seneca, and the satire of Juvenal. Though Livy stood out, Ciceronianism only found a real champion again in the rhetorician Quintilian.

Silver Age, AD 18–133. After the first flush of enthusiasm for Augustan ideals of national regeneration, literature paid the price of political patronage. It became subtly sterilized; and Ovid was but the first of many writers actually suppressed or inhibited by fear. Only Tacitus and Juvenal, writing under comparatively tolerant emperors, turned emotions pent up under Domitian's reign of terror into the driving force of great literature. Late Augustans such as Livy already sensed that Rome had passed its summit. Yet the title of Silver Age is not undeserved by a period that produced, in addition to Tacitus and Juvenal, the two Senecas, Lucan, Persius, the two Plinys, Quintilian, Petronius, Statius, Martial, and, of lesser stature, Manilius, Valerius Flaccus, Silius Italicus, and Suetonius.

Later writers. The decentralization of the empire under Hadrian and the Antonines weakened the Roman pride and passion for liberty. Romans began again to write in Greek as well as Latin. The "new sophistic" movement in Greece affected the "novel poets" such as Florus. An effete culture devoted itself to philology, archaism, and preciosity. After Juvenal, 250 years elapsed before Ausonius of Bordeaux (4th century AD) and the last of the true classics, Claudian (flourished about 400), appeared. The anonymous *Pervigilium Veneris* ("Vigil of Venus"), of uncertain date, presages the Middle Ages in its vitality and touch of stressed metre. Ausonius, though in the pagan literary tradition, was a Christian and contemporary with a truly original Christian poet, the Spaniard Prudentius. Henceforward, Christian literature overlaps pagan and generally surpasses it.

In prose these centuries have somewhat more to boast, though the greatest work by a Roman was written in Greek, the *Meditations* of the emperor Marcus Aurelius. *Elocutio novella*, a blend of archaisms and colloquial speech, is seen to best advantage in Apuleius (born about 125). Other writers of note were Aulus Gellius and Macrobius. The 4th century AD was the age of the grammarians and commentators, but in prose some of the most interesting work is again Christian.

The *Meditations* of Marcus Aurelius

THE GENRES

Comedy. Roman comedy was based on the New Comedy fashionable in Greece, whose classic representative was Menander. But whereas this was imitation of life to the Greeks, to the Romans it was escape to fantasy and literary convention. Livius' successor, Naevius, who developed this "drama in Greek cloak" (*fabula palliata*), may have been the first to introduce recitative and song, thereby increasing its unreality. But he slipped in details of Roman life and outspoken criticisms of powerful men. His imprisonment warned comedy off topical references, but the Roman audience became alert in applying ancient lines to modern situations and in demonstrating their feelings by appropriate clamour.

Unlike his predecessors, Plautus specialized, writing only comedy involving high spirits, oaths, linguistic play, slapstick humour, music, and skillful adaptation of rhythm to subject matter. Some of his plays can be thought of almost as comic opera. Part of the humour consisted in the sudden intrusion of Roman things into this conventional Greek world. "The Plautine in Plautus" consists in pervasive qualities rather than supposed innovations of plot or technique.

As Greek influence on Roman culture increased, Roman drama became more dependent on Greek models. Terence's comedy was very different from Plautus'. Singing almost disappeared from his plays, and recitative was less prominent. From Menander he learned to exhibit refinements of psychology and to construct ingenious plots; but he lacked comic force. His pride was refined language—the avoidance of vulgarity, obscurity, or slang. His characters were less differentiated in speech than those of Plautus, but they talk with an elegant charm. The society Terence portrayed was more sensitive than that of Plautine comedy; lovers tended to be loyal and sons obedient. His historical significance has been enhanced by the loss of nearly all of Menander's work.

Though often revived, plays modeled on Greek drama were rarely written after Terence. The Ciceronian was the great age of acting, and in 55 BC Pompey gave Rome a permanent theatre. Plays having an Italian setting came into vogue, their framework being Greek New Comedy but their subject Roman society. A native form of farce was also revived. Under Julius Caesar, this yielded in popularity to verse mime of Greek origin that was realistic, often obscene, and full of quotable apothegms. Finally, when mime gave rise to the dumb show of the *pantomimus* with choral accompaniment and when exotic spectacles had become the rage, Roman comedy faded out.

Tragedy. Livius introduced both Greek tragedy (*fabula crepidata*, "buskined") and comedy to Latin. He was followed by Naevius and Ennius, who loved Euripides. Pacuvius, probably a greater tragedian, liked Sophocles and heightened tragic diction even more than Ennius. His successor, Accius, was more rhetorical and impetuous. The fragments of these poets betoken grandeur in "the high Roman fashion," but they also have a certain ruggedness. They did not always deal in Greek mythology: occasionally they exploited Roman legend or even recent history. The Roman chorus, unlike the Greek, performed on stage and was inextricably involved in the action.

Classical tragedy was seldom composed after Accius, though its plays were constantly revived. Writing plays, once a function of slaves and freedmen, became a pastime of aristocratic dilettantes. Such writers had commonly no thought of production: post-Augustan drama was for reading. The extant tragedies of the younger Seneca probably were not written for public performance. They are melodramas of horror and violence, marked by sensational pseudo-realism and rhetorical cleverness. Characterization is crude, and philosophical moralizing obtrusive. Yet Seneca was a model for 16th- and early 17th-century tragedy, especially in France, and influenced English revenge tragedy.

Epic and epyllion. Livius' pioneering *Odyssey* was, to judge from the fragments, primitive, as was the *Bellum Punicum* of Naevius, important for Virgil because it began with the legendary origins of Carthage in Phoenicia and Rome in Troy. But Ennius' *Annales* soon followed. This compound of legendary origins and history was in Latin, in a transplanted metre, and by a poet who had imagination and a realization of the emergent greatness of Rome. In form his work must have been ill-balanced; he almost ignored the First Punic War in consideration of Naevius and became more detailed as he added books about his own times. But his great merit shines out from the fragments—nobility of ethos matched with nobility of language. On receptive spirits, such as Cicero, Lucretius, and Virgil, his influence was profound.

Little is known of the "strong epic" for which Virgil's friend Varius is renowned, but Virgil's *Aeneid* was certainly something new. Recent history would have been too particularized a theme. Instead, Virgil developed Naevius' version of Aeneas' pilgrimage from Troy to found Rome. The poem is in part an Odyssey of travel (with an interlude of love) followed by an Iliad of conquest, and in part a symbolic epic of contemporary Roman relevance. Aeneas has Homeric traits but also qualities that look forward to the character of the Roman hero of the future. His fault was to have lingered at Carthage. The command to leave the Carthaginian queen Dido shakes him ruthlessly out of the last great temptation to seek individual happiness. But it is only the vision of Rome's future greatness, seen when he visits Elysium, that kindles obedient acceptance into imaginative enthusiasm. It was just such a sacrifice of the individual that the Augustan ideal demanded. The second half of the poem represents the fusing in the crucible of war of the civilized graces of Troy with the manly virtues of Italy. The tempering of Roman culture by Italian hardiness was another part of the Augustan ideal. So was a revival of interest in ancient customs and religious observances, which Virgil could appropriately indulge. The verse throughout is superbly varied, musical, and rhetorical in the best sense.

Virgil's *Aeneid*

With his *Hecale*, Callimachus had inaugurated the short, carefully composed hexameter narrative (called epyllion by modern scholars) to replace grand epic. The *Hecale* had started a convention of insetting an independent story. Catullus inset the story of Ariadne on Naxos into that of the marriage of Peleus and Thetis, and the poem has a mannered, lyrical beauty. But the story of Aristaeus at the end of Virgil's *Georgics*, with that of Orpheus and Eurydice inset, shows what heights epyllion could attain.

Ovid's *Metamorphoses* is a nexus of some 50 epyllia with shorter episodes. He created a convincing imaginative world with a magical logic of its own. His continuous poem, meandering from the creation of the world to the apotheosis of Julius Caesar, is a great Baroque conception, executed in swift, clear hexameters. Its frequent irony and humour are striking. Thereafter epics proliferated. Statius' *Thebaid* and inchoate *Achilleid* and Valerius' *Argonautica* are justly less read now than they were. Lucan's unfinished *Pharsalia* has a more interesting subject, namely the struggle between Caesar and Pompey, whom he favours. He left out the gods. His brilliant rhetoric comes close to making the poem a success, but it is too strained and monochromatic.

Didactic poetry. Ennius essayed didactic poetry in his *Epicharmus*, a work on the nature of the physical universe. Lucretius' *De rerum natura* is an account of Epicurus' atomic theory of matter, its aim being to free men from superstition and the fear of death. Its combination of moral urgency, intellectual force, and precise observation of the physical world makes it one of the summits of classical literature.

This poem profoundly affected Virgil, but his poetic reaction was delayed for some 17 years; and the *Georgics*, though deeply influenced by Lucretius, were not truly didactic. Country-bred though he was, Virgil wrote for literary readers like himself, selecting whatever would contribute picturesque detail to his impressionistic picture of rural life. The *Georgics* portrayed the recently united land of Italy and taught that the idle Golden Age of the fourth *Eclogue* was a mirage: relentless work, introduced by a paternal Jupiter to sharpen men's wits, creates "the glory of the divine countryside." The compensation is the infinite variety of civilized life. Insofar as it had a political intention, it encouraged revival of an agriculture devastated in wars, of the old Italian virtues, and of the idea of Rome's extending its works over Italy and civilizing the world.

Ovid's *Ars amatoria* was comedy or satire in the burlesque guise of didactic, an amusing commentary on the psychology of love. The *Fasti* was didactic in popularizing the new calendar; but its object was clearly to entertain.

Satire. *Satura* meant a medley. The word was applied to variety performances introduced, according to Livy, by the Etruscans. Literary satire begins with Ennius, but it was Lucilius who established the genre. After experimenting, he settled on hexameters, thus making them its recognized vehicle. A tendency to break into dialogue may be a vestige of a dramatic element in nonliterary *satura*. Lucilius used this medium for self-expression, fearlessly criticizing public as well as private conduct. He owed much to the Cynic-Stoic "diatribes" (racy sermons in prose or verse) of Greeks such as Bion; but in extant Hellenistic literature he is most clearly presaged by the fragments of Callimachus' iambs. "Menippean" satire, which descended from the Greek prototype of Menippus of Gadara and mingled prose and verse, was introduced to Rome by Varro.

The satire of Horace and Juvenal

Horace saw that satire was still awaiting improvement: Lucilius had been an uncouth versifier. *Satires* I, 1–3 are essays in the Lucilian manner. But Horace's nature was to laugh, not to flay, and his incidental butts were either insignificant or dead. He came to appreciate that the real point about Lucilius was not his denunciations but his self-revelation. This encouraged him to talk about himself. In *Satires* II he developed in parts the satire of moral diatribe presaging Juvenal. His successor Persius blended Lucilius, Horace, diatribe, and mime into pungent sermons in verse. The great declaimer was Juvenal, who fixed the idea of satire for posterity. Gone was the personal approach of Lucilius and Horace. His anger may at times have been cultivated for effect, but his epigrammatic power and brilliant eye for detail make him a great poet.

The younger Seneca's *Apocolocyntosis* was a medley of prose and verse, but its pitiless skit on the deification of the emperor Claudius was Lucilian satire. The *Satyricon* of Petronius is also Menippean inasmuch as it contains varied digressions and occasional verse; essentially, however, it comes under fiction.

With Lucilian satire may be classed the fables of Augustus' freedman Phaedrus, the Roman Aesop, whose beast fables include contemporary allusions.

Iambic, lyric, and epigram. The short poems of Catullus were called by himself *nugae* ("trifles"). They vary remarkably in mood and intention, and he uses iambic metre normally associated with invective not only for his abuse of Caesar and Pompey but also for his tender homecoming to Sirmio. Catullus alone used the hendecasyllable, the metre of skits and lampoons, as a medium for love poetry.

Horace was a pioneer. In his *Epodes* he used iambic verse to express devotion to Maecenas and for brutal invective in the manner of the Greek poet Archilochus. But his primary aim was to create literature, whereas his models had been venting their feelings. In the *Odes* he adapted other Greek metres and claimed immortality for introducing early Greek lyric to Latin. The *Odes* rarely show the passion now associated with lyric but are marked by elegance, dignity, and studied perfectionism.

Martial went back to Catullus for his metres and his often obscene wit. He fixed the notion of epigram for posterity by making it characteristically pointed.

Elegy. The elegiac couplet of hexameter and pentameter (verse line of five feet) was taken over by Catullus, who broke with tradition by filling elegy with personal emotion. One of his most intense poems in this metre, about Lesbia, extends to 26 lines; another is a long poem of involved design in which the fabled love of Laodameia for Protesilaus is incidentally used as a paradigm. These two poems make him the inventor of the "subjective" love elegy dealing with the poet's own passion. Gallus, whose work is lost, established the genre; Tibullus and Propertius smoothed out the metre.

Catullus' love poems

Propertius' first book is still Catullan in that it seems genuinely inspired by his passion for Cynthia: the involvement of Tibullus is less certain. Later, Propertius grew more interested in manipulating literary conventions. Tibullus' elegy is constructed of sections of placid couplets with subtle transitions. These two poets established the convention of the "soft poet," valiant only in the campaigns of love, immortalized through them and the Muses. Propertius was at first impervious to Augustan ideals, glorying in his abject slavery to love and his naughtiness (*nequitia*), though later he became acclimatized to Maecenas' circle.

Tibullus, a lover of peace, country life, and old religious customs, had grace and quiet humour. Propertius, too, could be charming, but he was far more. He often wrote impetuously, straining language and associative sequence with passion or irony or sombre imagination.

Ovid's aim was not to unburden his soul but to entertain. In the *Amores* he is outrageous and amusing in the role adopted from Propertius, his Corinna being probably a fiction. Elegy became his characteristic medium. He carried the couplet of his predecessors to its logical extreme, characterized by parallelism, regular flow and ebb, and a neat wit.

OTHER LANGUAGE AND LITERARY ART FORMS

Rhetoric and oratory. Speaking in the forum and law courts was the essence of a public career at Rome and hence of educational practice. After the 2nd century BC, Greek art affected Latin oratory. The dominant style in Cicero's time was the "Asiatic"—emotional, rhythmical, and ornate. Cicero, Asiatic at first, early learned to tone down his style. Criticized later by the revivers of plain style, he insisted that style should vary with subject. But in public speaking he held that crowds were swayed less by argument than emotion. He was the acknowledged master speaker from 70 BC until his death (43 BC). He expounded the history of Roman oratory in the *Brutus* and his own methods in the *De oratore*.

The establishment of monarchy robbed eloquence of its public importance, but rhetoric remained the crown of education. Insofar as this taught boys to marshal material clearly and to express themselves cogently, it performed the function of the modern essay; but insofar as the temptations of applause made it strained and affected, it did harm.

In the *De oratore*, Cicero had pleaded that an orator's training should be in all liberal arts. Education without rhetoric was inconceivable; but what Cicero was proposing was to graft onto it a complete system of higher education. Quintilian, in his *Institutio oratoria*, went back to Cicero for inspiration as well as style. Much of that work is conventional, but the first and last books in particular show admirable common sense and humanity; and his work greatly influenced Renaissance education.

History. Quintus Fabius Pictor wrote his pioneering history of Rome during the Second Punic War, using public and private records and writing in Greek. His immediate successors followed suit. Latin historical writing began with Cato's *Origines*. After him there were as many historiasters, or worthless historians, as the poetasters disdained by Cicero. The first great exception is Caesar's *Commentaries*, a political apologia in the guise of unvarnished narrative. The style is dignified, terse, clear, and unrhetorical.

Sallust took Thucydides as his model. He interpreted, using speeches, and ascribed motives. In his extant monographs *Bellum Catilinae* and *Bellum Jugurthinum*, he displays a sardonic moralism, using history to emphasize the decadence of the dominant caste. The revolution in style he inaugurated gives him importance.

Livy began his 40 years' task as Augustus came to power. His work consummated the annalistic tradition. If in historical method he fell short of modern standards, he had the literary virtues of a historian. He could vividly describe past events and interpret the participants' views in eloquent speeches. He inherited from Cicero his literary conception of history, his copiousness, and his principle of accommodating style to subject. Indeed, he was perhaps the greatest of Latin stylists. His earlier books, where his imagination has freer play, are the most readable. In the later books, the more historical the times become, the more disturbing are his uncritical methods and his patriotic bias. Livy's work now is judged mainly as literature.

The histories of Tacitus

Tacitus, on the other hand, stands higher now than in antiquity. Though his anti-imperial bias in attributing motives is plain, his facts can rarely be impugned; and his evocation of the terrors of tyranny is unforgettable. He is read for his penetrating characterizations, his drama, his ironical epigrams, and his unpredictability. His is an extreme development of the Sallustian style, coloured with archaic and poetic words, with a careful avoidance of the commonplace.

Suetonian biography apart, historiography thereafter degenerated into handbooks and epitomes until Ammianus Marcellinus appeared. He was refreshingly detached, rather ornate in style, but capable of vivid narrative and description. He continued Tacitus' account from Domitian's death to AD 378, more than half his work dealing with his own times.

Biography and letters. The idea of comparing Romans with foreigners was taken up by Cornelius Nepos, a friend of Cicero and Catullus. Of his *De viris illustribus* all that survive are 24 hack pieces about worthies long dead and one of real merit about his friend Atticus. The very fact that Atticus and Tiro decided to publish nearly 1,000 of Cicero's letters is evidence of public interest in people. Admiration of these fascinating letters gave rise to letter writing as a literary genre. The younger Pliny's letters, anticipating publication, convey a possibly rose-tinted picture of civilized life. They are nothing to his spontaneous correspondence with Trajan, where one learns of routine problems, for instance with Christians confronting a provincial governor in Bithynia. The letter as a verse form, beginning with striking examples by Catullus, was established by Horace, whose *Epistles* carry still further the humane refinement of his gentler satires.

Suetonius' lives of the Caesars and of poets contain much valuable information, especially since he had access to the imperial archives. His method was to cite in categories whatever he found, favourable or hostile, and to leave this raw material to the judgment of the reader. The *Historia Augusta*, covering the emperors from 117 to 284, is a collection of lives in the Suetonian tradition. Tacitus' *Agricola* was an admiring, but not necessarily overcoloured, biographical study.

Some of the most valuable autobiography was incidental, such as Cicero's account of his oratorical career in the *Brutus*. Horace's largely autobiographical *Epistles* I was sealed with a miniature self-portrait. Ovid, in exile and afraid of fading from Rome's memory, gave an invaluable account of his life in *Tristia* IV.

Philosophical and learned writings. The practical Roman mind produced no original philosopher. Apart from Lucretius the only name that demands consideration is Cicero's. He was trained at Athens in the eclectic New Academy, and eclectic he apparently remained, seeking a philosophy to fit his own constitution rather than a logical system valid for all. He used the dialogue form, avowedly in order to make people think for themselves instead of following authority. Essentially, he was a philosophical journalist, composing works that became one of the means by which Greek thought was absorbed into early Christian thinking. The *De officiis* is a treatise on ethics. The dialogues do not follow the Platonic, or dialectic, pattern but the Aristotelian, in which speakers expounded already formed opinions at greater length.

Nor were the Romans any more original in science. Instead, they produced encyclopaedists such as Varro and Celsus. Pliny's *Natural History* is a fascinating ragbag, especially valuable for art history, though it shows to what extent Hellenistic achievement in science had become confused or lost.

Literary criticism. Cicero's *Brutus* and the 10th book of Quintilian's *Institutio oratoria* provide examples of general criticism. Cicero stressed the importance of a well-stocked mind and native wit against mere handbook technique. By Horace's day, however, it had become more timely to insist on the equal importance of art. Some of Horace's best criticism is in the *Satires* (I, 4 and 10; II, 1), in the epistle to Florus (II, 2), and in the epistle to Augustus (II, 1), a vindication of the Augustans against archaists. But it was his epistle to Piso and his sons (later called *Ars poetica*) that was so influential throughout Europe in the 18th century. It supported, among acceptable if trite theses, the dubious one that poetry is necessarily best when it mingles the useful (particularly moral) with the pleasing. Much of the work concerned itself with drama. The Romans were better at discussing literary trends than fundamental principles—there is much good sense about this in Quintilian, and Tacitus' *Dialogus* is an acute discussion of the decline of oratory.

Horace's *Ars poetica*

Fiction. Republican and early imperial Rome knew no Latin fiction beyond such things as Sisenna's translation of Aristides' *Milesian Tales*. But two considerable works have survived from imperial times. Of Petronius' *Satyricon*, a rambling picaresque novel, one long extract and some fragments remain. The disreputable characters have varied adventures and talk lively colloquial Latin. The description of the vulgar parvenu Trimalchio's banquet is justly famous. Apuleius' *Metamorphoses* (*The Golden Ass*) has a hero who has accidentally been changed into an ass. After strange adventures he is restored to human shape by the goddess Isis. Many passages, notably the story of Cupid and Psyche, have a beauty that culminates in the apparition of Isis and the initiation of the hero into her mysteries. (L.P.Wi./R.H.A.J.)

Medieval Latin literature

From about 500 to 1500 Latin was the principal language of the church, as well as of administration, theology, philosophy, science, history, biography, and belles lettres, and medieval Latin literature is therefore remarkably rich. Two themes dominate the linguistic and literary development of medieval Latin: its close and creative adaptation of the classical heritage from which it emerged and its changing relationship with the medieval vernacular languages. Within these two broad themes a number of subsidiary yet significant strains can be distinguished: the emergence of national characteristics in the Latin literature produced in different parts of Europe; the refinement of the polarity between popular and learned Latin by the clergy's use of a colloquialism intelligible to its audience as a lingua franca; and the effect of certain periods of special vigour and artistic self-awareness, such as the Carolingian revival of the 8th and 9th centuries and the new impulse given to learned and vernacular literature in the 12th.

THE 3RD TO THE 5TH CENTURY: THE RISE OF CHRISTIAN LATIN LITERATURE

The early history of medieval Latin literature is in part the story of the reception of the classical past by the Christians, to whom it represented secular culture. Old forms and genres were continuously renewed over the millennium following the entrance of Christians to the circle of literary production, dated for convenience to the conversion of Constantine to Christianity (about AD 313). For example, the Latin epic persisted in recognizable form throughout the period, and its authors remained in continuous contact with the great classical exponents Lucan, Statius, and, above all, Virgil. From the 4th century, the degree of scholarly interpretation applied to these epic po-

Persistence of the Latin epic

ets, especially Virgil, was intensified. Virgilian technique was imitated by many poets, among them the 4th-century Spaniard Juvencus, who versified a portion of the Bible, and the author of the epic poem *Waltharius* (probably 9th century), written in hexameters.

Even before the conversion of Constantine, Christians were developing new forms of literature, which persisted throughout the ensuing centuries. The production of hagiographical texts (lives of the saints) was widespread in the Middle Ages. The first Acts of the Martyrs in Latin were written during the 3rd century, and the flowering of the form after the end of the period of persecution of Christians shows the powerful appeal that it exercised at all levels of society. The *Passio Sanctarum Perpetuae et Felicitatis* (*The Passion of SS. Perpetua and Felicity*), written in a style that owes little to classical precedent, is a distinctive early example of the genre.

The 3rd and 4th centuries were above all an age of translation. Among the Greek patristic writings diffused to a wider audience in the West in Latin versions, the lives of the Desert Fathers occupied an important place. The Latin translation by Evagrius, bishop of Antioch, of Athanasius' *Life of Saint Antony* enjoyed the widest transmission, and its influence is as marked by contrast in the early Latin Lives of the Saints as it is by imitation. Sulpicius Severus' biography of St. Martin, an original Latin work, greatly influenced hagiography over many centuries. (A further, equally influential example of the genre was the *Dialogues* of Pope Gregory the Great, written in about 593.)

The most important work of translation appeared at the end of the 4th century: the Vulgate, completed by the monastic leader Jerome, replaced sporadic earlier attempts to render the Bible into Latin. The idiom and style of the Bible's original languages were apparent through the veil of Jerome's Latin, however, and provided a counterweight to the classical styles that continued to be taught and practiced through the schools in the West. Exegesis of the text occupied many of the greatest minds of the Middle Ages for the largest part of their careers, and the literary work of many major authors, from Augustine and Gregory to Bede, reflects their individual understanding of Scripture.

The early Christian liturgy also gave birth to new forms of literature. From the ancient practice of psalmody in the churches derives the hymn. Ambrose, bishop of Milan in the second half of the 4th century, wrote the earliest prosaic hymns, which incorporated nonliturgical texts into the mass to be sung by the congregation. These were rapidly imitated, notably by the Spanish poet Prudentius at the end of the century, and remained in continuous use in churches and monasteries for more than a millennium.

A major problem of Christian thinkers in these centuries was the integration of the history of the pagan empire with the history of salvation. Synthesis and epitome of biblical and classical history appeared in the *Historiarum adversus paganos libri VII* (*7 Books of Histories Against the Pagans*) of Orosius and the briefer *Chronica* (*c.* 402–404) of Sulpicius Severus. On a larger scale, Augustine's *De civitate Dei* (*The City of God*) offered a comprehensive view of past history, the present, and the world to come in the light of scriptural revelation. His spiritual autobiography, the *Confessiones* (*Confessions*), was an exploration of the philosophical and emotional development of an individual soul. The distinctive originality of this work owed little to classical autobiography and was unmatched by later imitations.

Writings of Augustine

The Gallic schools of the 5th century gave rise to a literary culture unique in this period. Versification of the Bible developed a new degree of exegetical and stylistic refinement, while the letters of Paulinus of Nola and Sidonius Apollinaris, bishop of Auvergne, display a picture of cultivated aristocratic and ecclesiastical society. Both men were also admired as poets, Sidonius in particular as an encomiast. On the secular side, at the beginning of the century in Rome the Egyptian poet Claudian produced the most elaborate examples of imperial verse panegyric to a succession of dignitaries. His *Raptus Proserpinae* (*c.* 400; *The Rape of Proserpine*) is one of the last examples of an extended narrative in verse that dwells wholly in the world of pagan mythology.

THE 6TH TO THE 8TH CENTURY

Gaul's literary history is interrupted by the Frankish invasions, though there are signs that abbots and bishops began to perceive the benefit of using literature to promote the cults of local saints. Two figures of note are Gregory of Tours and Venantius Fortunatus, bishop of Poitiers. In addition to a vast corpus of hagiography, Gregory produced the monumental *Historia Francorum* (605–664; *History of the Franks*), the most extensive history of a barbarian people that had yet been written. He set the arrival of the Franks in Gaul, and their recent past, in the perspective of universal history.

An element of local patriotism is also discernible in his writings. Gregory was one of the many patrons who inspired the poet Fortunatus, whose astute and pliable talent achieved distinction in both secular panegyric and hymnody. His hagiography, in verse and in prose, also is prominent. His style exercised a powerful appeal upon the poets of the Carolingian renaissance.

Three figures of encyclopaedic learning dominate the literature of the 6th and 7th centuries. In the course of his long retirement from a career in public service under the Ostrogothic kings in Italy, Cassiodorus combined zealous preservation of the literature of the classical past with an enormously influential educational plan. His late 6th-century compendium of sacred and secular learning, *Institutiones divinarum et humanarum lectionum* (*An Introduction to Divine and Human Readings*), was among the shaping influences upon monastic culture. The Roman Boethius, a Neoplatonist philosopher, wrote on arithmetic and music, but his most popular and influential work was *De consolatione philosophiae* (1882–91; *The Consolation of Philosophy*), written in about 524, when Boethius was imprisoned under sentence of execution. The Spaniard Isidore produced a series of encyclopaedic compilations that were used as repositories of diverse learning by later centuries. It was midway through the 6th century that the last major Latin work was produced in the Eastern Empire: the epic *Iohannis* of the African poet Corippus.

The conversion of the Saxons began to bear literary fruit during the 7th and early 8th centuries. In an elaborate and allusive style, Aldhelm, bishop of Sherborne, wrote, first in prose and later in verse, a treatise on sainthood called *De Virginitate.* In the kingdom of Northumbria, particularly open to influence of Irish monastic learning, St. Bede the Venerable devoted his life to scholarship. The culmination of his work is the *Historia ecclesiastica gentis Anglorum* (*The Venerable Bede's Ecclesiastical History of England*), completed in 731. Synthesized from a variety of sources, literary and nonliterary, the work charts the involvement of God with the English people and the relation of the English church to the Christian world centred on Rome.

Writings of the Venerable Bede

THE CAROLINGIAN RENAISSANCE

The revival of letters, accompanied by wide-scale copying of classical texts, to which the reign of Charlemagne (768–814) gave fresh impetus, produced some of the most brilliant literary achievements of the Latin Middle Ages. An international elite of scholars, among whom the most distinguished were the Anglo-Saxon Alcuin, the Visigoth Theodulf of Orléans, and the Italians Paulinus of Aquileia and Paul the Deacon, produced a body of lyric, epic, and didactic poetry (both sacred and secular, both religious and political) unmatched in the earlier period. The revival of epic, and the secularization of the sacred hero, occurred in the extant third book of a lost and larger Virgilian epic, anonymously transmitted but known by the title *Karolus Magnus et Leo Papa* ("Charlemagne and Pope Leo"). Its example was followed in the next generation by Ermoldus Nigellus, writing about the deeds of Louis the Pious, and the tradition of earlier Carolingian authors is extended by two major political poets, Walafrid Strabo and Sedulius Scottus (also the author of an uproarious mock epyllion). In prose the major achievements lie in the fields of biography, with Einhard's *Vita Karoli Magni* (*c.* 830; *Life of Charlemagne*); of religious controversy, with Theodulf's *Libri Carolini* (defenses written at Charlemagne's request); and of theology, with John Duns Scotus' metaphysical masterpiece, the *Periphyseon.*

THE 9TH TO THE 11TH CENTURY

The sequence and the liturgical drama

From the later 9th century on, the liturgy gave rise to two new literary forms: the sequence and the liturgical drama. Notker Balbulus, monk of St. Gall, was not the first to compose sequences, but his *Liber hymnorum* ("Book of Hymns"), begun about 860, is an integrated collection of texts that spans the whole of the church year in an ordered cycle. Performed between the biblical readings in the mass, each sequence is a free meditation upon scriptural themes, often drawing upon and synthesizing disparate texts. Among later exponents of the genre, Adam of St. Victor was the most distinguished, though the mystical sequences of Hildegard of Bingen exercise a potent appeal. During the same period the enormous expansion of the cult of the Virgin left a notable mark upon hymnody, the early 11th century seeing the composition of Marian hymns, including such ubiquitous texts as "Salve Regina" ("Hail, Queen") and "Alma Redemptoris Mater" ("Sweet Mother of the Redeemer").

Notker's sequences are alive with dramatic possibility, and at St. Gall the practice of troping, or embellishing, liturgical texts also took dramatic form. The *Quem quaeritis* trope from St. Martial, an abbey at Limoges, was one of the earliest such pieces to demand dramatic performance. From this beginning developed the long tradition of liturgical drama, which, like the sequence, is centred upon the major feasts of the church year.

Two narrative works stand out in this period. The *Waltharius* epic is set in the years of the invasions of Attila the Hun. The sophistication of its narrative technique contrasts with its Germanic subject matter. The *Ruodlieb,* a romance written perhaps in about 1050 in a language heavily influenced by vernacular usage, reveals a comparable narrative subtlety. Even in its fragmentary state, the variety and vigour of its episodes are apparent.

The ease with which religious forms such as the sequence are adapted for secular use is nowhere seen better than in the 11th-century compilation known as the *Cambridge Songs.* The blend of humorous contes, hymnody, and lyric testifies to a diverse taste in the unknown anthologist. Other lyric collections from the next century, such as the Ripoll and Arundel lyrics, may draw upon work of earlier provenance. To the chance survival of individual compilations such as these derives the bulk of knowledge of the secular lyric, which is one of the chief distinctions of the 12th and 13th centuries.

THE 12TH TO THE 14TH CENTURY

The *Carmina Burana*

The *Carmina Burana* ("Songs from Bavaria"), the largest and greatest collection of secular lyrics, comes from the Benediktbeuern, a Benedictine monastery in Bavaria. It was put together in the 13th century, though most of the songs are much older, and contains work by many of the finest poets of the age. The contents are divided by subject into moral and satirical verse, love poetry, drinking songs, and liturgical dramas. Walter of Châtillon and Philip the Chancellor are conspicuous among the authors of the satires, the force of their works deriving from learned and allusive use of Scripture. Peter of Blois is found in the section of satirical verse and the section of love poetry. His verse forms achieve a new degree of delicacy and sophistication, and his erotic poetry owes much to a close study of classical poets, particularly Ovid. Yet many of the forms in evidence, the pastourelle (a love debate between a knight and a shepherdess) for example, have no classical antecedent. In the complexity of its argument and profusion of imagery, a poem such as "Dum Diane vitrea" ("While Shining Diane") far exceeds the imagination of any classical author. Among the drinking songs in the third section are works of the anonymous German "Archpoet" and of Hugh Primas of Orléans, a slightly earlier figure. Under the cover of a pointedly low-life persona, these poets, both prominent men in court society, practiced a robust form of satire in which much of the humour is deflected upon themselves. Grander forms of poetry are not neglected: Walter of Châtillon's foray into epic, the *Alexandreis* (written *c.* 1180), is one of the most distinguished products of the medieval fascination with the legends of Alexander the Great, and it exercised an immense influence on subsequent vernacular literature.

The 12th century was an age of philosophical development, above all in the cathedral schools (as at Chartres) and new universities (as at Paris). Scholars such as Alain of Lille (Alanus de Insulis) and John of Salisbury returned to philosophical problems that had been posed in the days of Boethius. With Roger Bacon, Duns Scotus, and Robert Grosseteste, the first chancellor of Oxford University, a significant English contribution is discernible. Peter Abelard trained at Paris, where he taught John of Salisbury. Of Abelard's philosophical works, *Sic et non* (completed *c.* 1136; "Yes and No") is the most notable, probing critically the vast bulk of received authority. In three of his most original literary works, the relationship with Héloïse is a prominent feature. The *Hymnarius Paraclitensis* is a collection of hymns for Héloïse's convent, where the reading of Scripture is complex and shows the imprint of novel theological thought. The six *planctus* ("laments") are meditations on guilt and suffering, set in the mouths of biblical personages, while the correspondence between Abelard and Héloïse reflects themes found in both verse collections. Abelard's autobiographical work, the *Historia calamitatum* (written *c.* 1136; *The Story of Abelard's Adversities*), recounts the story of his tragic love affair and its theological consequences.

Works of Peter Abelard

Liturgical and cultic innovation left its mark upon Latin literature during the 13th and 14th centuries. John of Garland's compilation of hymns to the Virgin is a late testimony to the force of Marian inspiration. From the early 13th century derive two of the latest sequences to feature in the liturgy in all countries, the "Dies irae" ("The Day of Wrath") and the "Stabat Mater" ("The Mother Stands"). The cults of the Holy Cross and of the Passion are the impetus to the poetry of two Franciscans, the Italian St. Bonaventura and John Pecham in England. Pecham's *Philomena praevia* is an extended lyrical meditation that blends the story of the Redemption with the liturgical course of a single day.

The theology of the 13th century is dominated in bulk and stature by the writings of St. Thomas Aquinas. The culmination of a career centred upon Paris and Rome is the *Summa theologiae* (written between 1265 and 1272), a systematic exposition of the essentials of faith, grounded in Aristotelian principles. The translation of Aristotle into Latin continued throughout the century. Aquinas' liturgical works also remained prevalent. (P.Go.)

Renaissance Latin literature

The term Renaissance Latin is associated, for 14th-century Italy, mainly with Dante, Petrarch, and Boccaccio, though mention should also be made of the Florentine historian Leonardo Bruni and the humanist scholars Albertino Mussato, Coluccio Salutati, and Aeneas Silvius Piccolomini (Pope Pius II). In verse there was a general return to classical models and elegance, while in prose Latin was still a necessary medium for the abundant humanistic, scientific, philosophical, and religious literature that was a mark of the new age.

In Italy there were three main centres of learning and literature in the 15th and 16th centuries: Florence, Rome, and Naples. Each had its own circle of writers and scholars. The Florentine group was noted for the Platonist philosophers Poggio Bracciolini, Marsilio Ficino, Giovanni Pico della Mirandola, and a poet and scholar, Angelo Poliziano. Rome was the centre for a grammarian, Pietro Bembo, and for Marco Vida, author of a Latin epic on the redemption, while Naples was the home of poets and scholars, notably Giovanni Pontano, Jacopo Sannazzaro, Lorenzo Valla, and Girolamo Fracastoro.

Influence of the Brethren of the Common Life

Germany and the Low Countries also made a large contribution in prose and verse to Latin literature in the 15th and 16th centuries. Many humanists owed their early education to the Brethren of the Common Life, a Dutch Christian community that laid great emphasis on the classics. Among these was Desiderius Erasmus, the greatest figure of the northern Renaissance. Bred in the rhetorical tradition of literary humanism, he had little interest in the scientific premonitions of the age. As an editor and expositor of classical texts and the writings of the Church Fathers, as a commentator on the ecclesiastical conflicts of his time, and as a scholar, wit, and satirist, he was unsurpassed by any humanist in northern Europe. A German abbot, Johannes Trithemius, was a historian and scholar with an immense range of interests and knowledge; Conradus Celtis was conspicuous as a humanist and poet; while Petrus Lotichius wrote elegant verse.

Spanish humanism was best seen in the scholar and friend of Erasmus, Juan Vives, while in England the statesman and scholar Sir Thomas More was the outstanding figure. Polydore Virgil, an Italian, brought the new methods of historical writing into England, though a poet and historian, Tito Livio Frulovisi, had written a life of Henry V that influenced later English writers. Among many Latin poets should be mentioned George Buchanan and John Barclay, both Scots. The strong English tradition of classical verse composition in the schools was shown in the Latin poems of such 17th-century poets as John Milton, Henry Vaughan, Richard Crashaw, and Abraham Cowley.

In France, where, as in England, the Renaissance came late, some members of the group of writers known as La Pléiade wrote Latin verse. Despite the eventual triumph of the French vernacular, Latin poems continued to be written, and several hymns composed in classical forms were included in church services in the 17th and 18th centuries. (F.J.E.R.)

Until the early 18th century, Latin was recognized as the best medium for historical and scientific work if it were intended to reach a European audience. For this reason Marsilio Ficino and Pico della Mirandola, Erasmus and More, and later Francis Bacon, Hugo Grotius, René Descartes, Benedict Spinoza, and Sir Isaac Newton used what was still an international language.

BIBLIOGRAPHY

Ancient: Detailed and documented accounts include W.S. TEUFFEL, *Teuffel's History of Roman Literature,* new ed., 2 vol. (1891–92, reprinted 1967; originally published in German, 5th rev. ed. 1890); MARTIN SCHANZ, *Geschichte der römischen Literatur bis zum Gesetzgebungswerk des Kaisers Justinian,* rev. by CARL HOSIUS and G. KRUGER, 4 vol. in 5 (1914–35, reissued 1966–71); AUGUSTO ROSTAGNI, *Storia della letterature latina,* 3rd ed., 3 vol. (1964), sumptuously illustrated; EDUARD NORDEN, *Die römische Literature,* 6th ed. (1961); H.J. ROSE, *A Handbook of Latin Literature from the Earliest Times to the Death of St. Augustine,* 3rd ed. (1954, reprinted with supp. bibliog., 1966); and E.J. KENNEY and W.V. CLAUSEN (eds.), *The Cambridge History of Classical Literature,* vol. 2, *Latin Literature* (1982). J. WIGHT DUFF, *A Literary History of Rome: From the Origins to the Close of the Golden Age,* 3rd ed., ed. by A.M. DUFF (1953, reprinted 1967), and *A Literary History of Rome in the Silver Age: From Tiberius to Hadrian,* 3rd ed., ed. by A.M. DUFF (1964), is a standard introduction containing comprehensive and scholarly surveys, with supplementary bibliographies. Shorter accounts are J.W. MACKAIL, *Latin Literature,* 2nd rev. ed. (1896, reissued 1966); MICHAEL GRANT, *Roman Literature,* new ed. (1958); and KARL BUCHNER, *Römische Literaturgeschichte,* 5th ed. (1980). On poets, see W.Y. SELLAR, *The Roman Poets of the Republic,* 3rd ed. (1889, reprinted 1965), *The Roman Poets of the Augustan Age: Virgil,* 3rd ed. (1897, reprinted 1965), and *The Roman Poets of the Augustan Age: Horace and the Elegiac Poets,* 2nd ed. (1899, reprinted 1965); H.E. BUTLER, *Post-Augustan Poetry from Seneca to Juvenal* (1909, reprinted 1977); and J. WIGHT DUFF, *Roman Satire* (1936, reissued 1964). General works include H. BARDON, *La Littérature latine inconnue* (1952); F. KLINGNER, *Römische Geisteswelt,* 5th ed. (1965, reprinted 1979); M.L. CLARKE, *Rhetoric at Rome* (1953, reprinted 1968), and *The Roman Mind* (1956, reissued 1968); S.F. BONNER, *Roman Declamation in the Late Republic and Early Empire* (1949, reprinted 1969); W. BEARE, *The Roman Stage,* 3rd rev. ed. (1964, reprinted 1977), and *Latin Verse and European Song* (1957); and GORDON WILLIAMS, *Tradition and Originality in Roman Poetry* (1968). For an account of Latin studies see M. PLATNAUER (ed.), *Fifty Years (and Twelve) of Classical Scholarship,* 2nd rev. ed. (1968); and the annual bibliography in *L'Année philologique,* published by the International Society of Classical Bibliography in Paris. Posthumous influence of various authors is traced in G. HIGHET, *The Classical Tradition* (1949, reprinted 1957); and R.R. BOLGAR, *The Classical Heritage and Its Beneficiaries* (1954, reprinted 1977).

Middle Ages: ADOLF EBERT, *Allgemeine Geschichte der Literatur des Mittelalters in Abendland,* 2nd ed., 3 vol. (1880–89, reprinted 1971); GUSTAVO VINAY, *Alto midioevo latino: conversazioni e no* (1978); BERNHARD BISCHOFF, *Mittelalterliche Studien,* 2 vol. (1966–67); CHRISTINE MOHRMANN, *Études sur le latin des chrétiens,* 3 vol. (1958–61); K. STRECKER, *Introduction to Medieval Latin* (1957, reissued 1968; originally published in German, 2nd ed., 1929); GIOVANNI CREMASCHI, *Guida allo studio del latino medievale* (1959); F.A. WRIGHT and T.A. SINCLAIR, *A History of Later Latin Literature from the Middle of the Fourth to the End of the Seventeenth Century* (1931, reprinted 1969); M. HÉLIN, *A History of Medieval Latin Literature,* rev. ed. (1949, originally published in French, 1943); PIERRE DE LABRIOLLE, *Histoire de la littérature latine chrétienne,* 2 vol., 3rd ed. (1947); M.L.W. LAISTNER, *Thought and Letters in Western Europe, A.D. 500 to 900,* new ed. (1957); MAX MANITIUS, *Geschichte der lateinischen Literatur des Mittelalters,* 3 vol. (1923–31, reprinted 1965–74); J. DE GHELLINCK, *Littérature latine au moyen age,* 2 vol. (1939), *L'Essor de la littérature latine au XII^e,* 2 vol. (1946); C.H. HASKINS, *The Renaissance of the Twelfth Century* (1927, reissued 1971); and ERNST ROBERT CURTIUS, *European Literature and the Latin Middle Ages* (1953, reprinted 1983; originally published in German, 1948). On the poetry of the period, see FREDERIC J.E. RABY, *A History of Christian-Latin Poetry from the Beginnings to the Close of the Middle Ages,* 2nd ed. (1953, reprinted 1966), and *A History of Secular Latin Poetry in the Middle Ages,* 2 vol., 2nd ed. (1957, reprinted 1967); MAX MANITIUS, *Geschichte der christlich-lateinischen Poesie* (1891); DAG L. NORBERG, *La Poésie latine rhythmique du Haut Moyen Age* (1954); JACQUES FONTAINE, *Naissance de la poésie dans l'occident Chrétien* (1981); and PETER GODMAN (ed.), *Poetry of the Carolingian Renaissance* (1985).

Renaissance: P. VAN TIEGHEM, *La Littérature latine de la Renaissance* (1944, reprinted 1966); WILFRED P. MUSTARD (ed.), *Studies in the Renaissance Pastoral,* 6 vol. (1911–31); GEORG ELLINGER, *Geschichte der neulateinischen Literatur Deutschlands im sechzehnten Jahrhundert,* 3 vol. (1929–33, reprinted 1969); WOLFGANG MANN, *Lateinische Dichtung in England vom Ausgang der Frühhumanismus bis zum Regierungsantritt Elisabeths* (1939); JOHN SPARROW, "Latin Verse of the High Renaissance," in E.F. JACOB (ed.), *Italian Renaissance Studies* (1960); ALESSANDRO PEROSA and JOHN SPARROW (eds. and comps.), *Renaissance Latin Verse: An Anthology* (1979); ROBERTO WEISS, *The Dawn of Humanism in Italy* (1947, reprinted 1970), *The Spread of Italian Humanism* (1964), and *Humanism in England During the Fifteenth Century,* 3rd ed. (1967); and HANS BARON, *The Crisis of the Early Italian Renaissance* (1966).

(Ed.)

Lavoisier

A French chemist and the father of modern chemistry, Antoine-Laurent Lavoisier was a brilliant experimenter and many-sided genius who was active in public affairs as well as in science. He developed a new theory of combustion that led to the overthrow of the phlogistic doctrine, which had dominated the course of chemistry for more than a century. His fundamental studies on oxidation demonstrated the role of oxygen in chemical processes and showed quantitatively the similarity between oxidation and respiration. He formulated the principle of the conservation of matter in chemical reactions. He clarified the distinction between elements and compounds and was instrumental in devising the modern system of chemical nomenclature. Lavoisier was one of the first scientific workers to introduce quantitative procedures into chemical investigations. His experimental ingenuity, exact methods, and cogent reasoning, no less than his discoveries, revolutionized chemistry. His name is indissolubly linked to the establishment of the foundations upon which modern science rests.

Education and early work

Lavoisier was born in Paris on August 26, 1743. His father, an *avocat au parlement* (parliamentary counsel), gave him an excellent education at the Collège Mazarin, where, along with a solid classical grounding in language, literature, and philosophy, he received the best available training in the sciences, including mathematics, astronomy, chemistry, and botany. Following his family's tradition, he pursued the study of law, and he received his license to practice in 1764. His inquiring mind, however, continually drew him to science. In 1766 he received a gold medal from the Academy of Sciences for an essay on the best means of lighting a large town. Among his early work were papers on the Aurora Borealis, on thunder, and on the composition of gypsum. Pursuing an early interest in rocks and minerals, he accompanied the geologist J.-E. Guettard on a long geological trip and assisted him in preparing his mineralogical atlas of France. In 1768, after presenting a paper on the analysis of water samples, Lavoisier was admitted to the academy as *adjoint-chimiste* (associate chemist). He passed through all the grades in the academic structure and was made director in 1785 and treasurer in 1791.

Through his family, Lavoisier became independently wealthy in his early 20s. In 1771 he married Marie Paulze, who would later assist him in his work by illustrating his experiments, recording results, and translating scientific articles from English. In accordance with a common practice among the wealthy bourgeoisie at the time, his father bought him a title of nobility in 1772, and a few years later Lavoisier purchased the country estate of Fréchines, near Blois.

Scientific achievements. Lavoisier's name gained wide recognition when, in 1770, he refuted the then prevalent belief that water is converted into earth by repeated distillation. By carefully weighing both the earthy residue and the distilling apparatus, he demonstrated that the solid matter came from the glass vessels and not from the water.

Study of combustion

Speculating on the nature of the traditional four elements—earth, water, air, and fire—Lavoisier began to investigate the role of air in combustion. On November 1, 1772, he deposited with the Academy of Sciences a note stating that sulfur and phosphorus when burned increased in weight because they absorbed "air," while the metallic lead formed when litharge was heated with charcoal weighed less than the original litharge because it had lost "air." The exact nature of the airs concerned in the processes he could not yet explain, and he proceeded to study the question extensively. In 1774 he published his first book, *Opuscules physiques et chimiques*, in which he presented the results of both his reading and his experimentation. That year Joseph Priestley prepared "dephlogisticated air" (oxygen) by heating "red precipitate of mercury." Lavoisier confirmed and extended Priestley's work. Perceiving that in combustion and the calcination of metals only a portion of a given volume of common air was used up, he concluded that the active agent was Priestley's new "air," which was absorbed by burning, and that "nonvital air," or azote (nitrogen), remained behind. He observed that birds lived longer in the new "eminently respirable air," as he described it, and he showed that this air combined with carbon to produce the "fixed air" (carbon dioxide) obtained by Joseph Black in 1754.

Recognition that the atmosphere is composed of different gases that take part in chemical reactions made it possible to identify the composition of many substances, particularly the acids. In a memoir presented to the academy in 1777, read in 1779 but not published until 1781, Lavoisier assigned to dephlogisticated air the name oxygen, or "acid producer," on the erroneous supposition that all acids were formed by its union with a simple, usually nonmetallic body. He explained combustion not as the result of the liberation of a hypothetical fire principle, phlogiston, but as the result of the combination of the burning substance with oxygen. On June 25, 1783, he announced to the academy that water was the product formed by the combination of hydrogen and oxygen; in this, however, he had been anticipated by the English chemist Henry Cavendish. As a member of a committee for finding ways to improve lighter-than-air flight with the newly invented balloons, he produced quantities of hydrogen, called "inflammable air," by decomposing water into its constituent gases. From his knowledge of the composition of water, Lavoisier was led to the beginnings of quantitative organic analysis. He burned alcohol and other combustible organic compounds in oxygen, and from the weight of water and carbon dioxide produced he calculated their composition.

Lavoisier published a brilliant attack on the phlogistic theory in 1786. Despite the opposition of Priestley and others, a growing number of scientists began to adopt his

By courtesy of The Rockefeller University, New York City

Lavoisier, with his wife, oil painting by Jacques-Louis David, 1788. In the collection of Rockefeller University, New York City.

views. In 1787 a group of French chemists published the *Méthode de nomenclature chimique,* which classified and renamed the known elements and compounds. Reflecting Lavoisier's new discoveries and theories, the *Nomenclature* exerted a wide influence. Also influential was the revision in 1788 of Antoine-François de Fourcroy's popular *Élémens d'histoire naturelle et de chimie,* which was completely recast in terms of Lavoisier's views and according to the new chemical nomenclature. The following year Lavoisier and others established the *Annales de chimie,* a journal devoted to the new chemistry. Gradually the older approach based on the phlogistic theory lost adherents, and eventually Lavoisier's ideas were adopted universally.

The spread of Lavoisier's doctrines was greatly facilitated by the defined and logical form in which he presented them in his *Traité élémentaire de chimie* (1789). This classic book provided a concise exposition of his work and that of his followers and offered an introduction to the new approach to chemistry. In the prefatory "Discours préliminaire" Lavoisier set forth his views on the proper methods of scientific inquiry and scientific teaching, and he defended the new nomenclature. Those substances that could not be decomposed he termed *substances simples,* the elements out of which other matter was made. To a large extent the modern concept of an element, as against the ancient Greek idea, stems from Lavoisier. In the *Traité* he furnished a clear statement of his principle of the conservation of matter in chemical reactions. Nothing, he said, is created or destroyed; there are only alterations and modifications, and there is an equal quantity—an equation—of matter before and after the operation.

Other scientific work

In addition to his purely chemical work, Lavoisier, mostly in conjunction with the mathematician and astronomer Pierre-Simon Laplace, devoted considerable attention to physical problems, especially those connected with heat. The two carried out some of the earliest thermochemical investigations, devised an apparatus for measuring linear and cubical expansions, and employed a modification of Black's ice calorimeter in a series of determinations of specific heats. Regarding heat (*matière du feu*) as a peculiar kind of imponderable matter, Lavoisier held that the three states of aggregation—solid, liquid, and gas—were modes of matter, each depending on the amount of *matière du feu* with which the substances concerned were associated. He also worked at fermentation, respiration, and animal heat, looking upon the processes concerned as essentially chemical in nature. From measurements made in his pioneering biochemical experiments on animal heat and on the gases exchanged during respiration, he concluded that respiration was a type of oxidation reaction similar to the burning of carbon. A paper discovered many years after his death showed that he had anticipated later thinkers in explaining the cyclical process of animal and vegetable life.

Public service. Throughout Lavoisier's extraordinary career as a scientist, he carried on a simultaneous career as a public servant of remarkable versatility, contributing his talents in the areas of finance, economics, agriculture, education, and social welfare, among others. In 1768 he became an assistant in one of the revenue-collecting departments of the government, subsequently becoming a full titular member of the Ferme Générale, the main tax-collecting agency. The financial and organizational abilities he displayed as a farmer-general, along with his undoubted scientific and technical capacity, led in 1775 to his appointment as *régisseur des poudres* (a director of the gunpowder administration). With his customary energy he set about making improvements in the chaotic powder industry. He abolished the vexatious search for saltpetre in the cellars of private houses, increased the production of the salt, and improved the manufacture of gunpowder. The post enabled Lavoisier to move to the Arsenal of Paris, where he took up residence and equipped a superb laboratory. This establishment soon became a gathering place for the scientists and advanced thinkers of the day, and the dinners presided over by his wife became famous. After dinner the guests often would be escorted to the laboratory to witness or take part in a demonstration of some new experiment. Although an increasing number of public duties claimed Lavoisier's time, he was able to regularly set aside one day a week for scientific investigations.

As his influence in the Academy of Sciences grew, so did his responsibilities. He was a member of numerous official committees to look into matters concerning the public. In 1781 the notorious Franz Anton Mesmer arrived in Paris, and Lavoisier (along with Benjamin Franklin) served on a committee to investigate his cures by "animal magnetism," pronouncing them a hoax. With another committee, he explored the hospitals and prisons of Paris and recommended remedies for their deplorable state. At Fréchines he started a model farm, where he demonstrated the advantages of scientific agriculture. In 1785 he was named to the government's committee on agriculture and as its secretary drew up reports and instructions on the cultivation of crops, promulgating various agricultural schemes. As a landowner in the province of Orléans, Lavoisier was chosen a member of the provincial assembly in 1787. There he devised ways to improve local social and economic conditions by means of savings banks, insurance societies, canals, workhouses, and tax reforms. He advanced money without interest to the towns of Blois and Romorantin for the purchase of barley during the famine of 1788. He was associated with committees on hygiene, coinage, the casting of cannon, and public education. He was secretary and treasurer of the commission appointed in 1790 to secure uniformity of weights and measures throughout France, work that led to the establishment of the metric system.

A reformer and political liberal, Lavoisier was active in the French Revolution. When the Estates-General was reconvened in 1789, he became an alternate deputy and drew up the code of instructions for guidance of the deputies. He was elected to the commune of Paris and joined the moderate Society of 1789, a planning group. As an administrator of the national treasury, he published analyses of France's finances and agriculture. But his membership in the unpopular Ferme Générale was alone sufficient to make him an object of suspicion to the authorities, and, despite his many services to the nation and his renown as a scientist, he came under increasing attack from radical pamphleteers. In 1787, at Lavoisier's suggestion, a wall had been erected around Paris to halt the influx of contraband. The extremist revolutionary Jean-Paul Marat accused him of imprisoning Paris and of stopping the circulation of air. In 1791 the Ferme Générale was abolished, and Lavoisier subsequently lost his position in the gunpowder administration and was forced to leave the Arsenal. In 1793 the Reign of Terror commenced. Despite strenuous efforts by Lavoisier, the Academy of Sciences, along with the other learned societies, was suppressed. At year's end the Revolutionary Convention ordered the arrest of the former members of the Ferme Générale, and on May 8, 1794, after a trial that lasted less than a day, a revolutionary tribunal condemned Lavoisier and 27 others to death. That afternoon, he and his companions, including his father-in-law, were guillotined at the Place de la Révolution (now Concorde). His body was thrown into a common grave.

Lavoisier's fate in the Revolution

BIBLIOGRAPHY. A definitive annotated survey of Lavoisier's writing is presented in DENIS I. DUVEEN and HERBERT S. KLICKSTEIN, *A Bibliography of the Works of Antoine Laurent Lavoisier, 1743–1794* (1955), with a *Supplement* (1965). ANTOINE LAURENT LAVOISIER, *Oeuvres de Lavoisier,* 6 vol. (1862–93, reissued 1965), is a collected edition that includes all his principal works and memoirs from the academy volumes, as well as numerous miscellaneous writings.

DOUGLAS MCKIE, *Antoine Lavoisier: Scientist, Economist, Social Reformer* (1952, reprinted 1980), is an authoritative biography. HENRY GUERLAC, *Lavoisier—The Crucial Year: The Background and Origin of His First Experiments on Combustion in 1772* (1961, reprinted 1990), and *Antoine-Laurent Lavoisier, Chemist and Revolutionary* (1975), focus on his research. Lavoisier's place in the history of science and in the development of scientific disciplines as independent branches of knowledge is explored in KENNETH S. DAVIS, *The Cautionary Scientists: Priestley, Lavoisier, and the Founding of Modern Chemistry* (1966); FREDERICK LAWRENCE HOLMES, *Lavoisier and the Chemistry of Life: An Exploration of Scientific Creativity* (1985); MARCO BERETTA, *The Enlightenment of Matter: The Definition of Chemistry from Agricola to Lavoisier* (1993); and ARTHUR DONOVAN, *Antoine Lavoisier: Science, Administration, and Revolution* (1993). (D.I.D./Ed.)

The Profession and Practice of Law

The primary function of the profession and practice of law is to apply the law in specific cases—to individualize it. This function is manifest in the work of the advocate and the judge in the process of trying and deciding cases. In Anglo-American systems a lawyer investigates the facts and the evidence by conferring with his client, interviewing witnesses, and reviewing documents. He may seek a summary dismissal because the opponent evidently has no case, or, through discovery proceedings he may force the other side to reveal more fully the issues and facts on which it relies. At the trial he introduces evidence, objects to improper evidence from the other side, and advances partisan positions on questions of law and of fact. In continental European countries the judge has greater responsibility for investigation of the facts. At trial he plays an active role in taking evidence, questioning witnesses, and framing the issues. Continental lawyers suggest lines of factual inquiry to the judge and, like their Anglo-American counterparts, advance legal theories and argue the law in accord with the interests of their clients. In either system, if a lawyer loses his client's case, he may seek a new trial or relief in an appellate court.

Even controversies that are not resolved in court require the aid of lawyers. Negotiation, reconciliation, compromise—in all of which lawyers have a large part—bring about the settlement of most cases without trial.

The profession also applies and utilizes the law in the less dramatic setting of the office. The lawyer as counselor and negotiator may aid in shaping a transaction so as to avoid disputes or legal difficulties in the future, or so as to achieve advantages for his client, such as the minimization of taxes. The law gives to private persons extensive but not unlimited power to arrange and determine their legal rights in many matters and in various ways, such as through wills, contracts, leases, or corporate bylaws. In structuring these arrangements the lawyer is helping to particularize the legal rights of the parties.

Another field of legal work, which has developed rapidly in the 20th century, is the representation of clients before administrative commissions, commissions of inquiry, and, in some countries, legislative committees. This development has been a result of the increase of government regulation of economic life.

A lawyer has several loyalties in his work, including loyalty to his client, to the administration of justice, to the community, to his associates in practice, and to himself—whether to his economic interests or to his ethical standards. These diverse and at times competing loyalties must be reconciled with wisdom. It is the purpose of the standards of the profession to effect the reconciliation.

(M.A.Gl.)

This article is divided into the following sections:

Legal profession

One definition of the legal profession is "the vocation based on expertness in the law and its application." This simple definition may be best, despite the fact that in some countries there are several professions and even some occupations (*e.g.*, police service) that require this expertness but may not be considered to be within the "legal profession" at all.

HISTORY

Distinct legal systems emerged relatively early in history, but legal professions of size and importance are relatively modern. There is not the slightest trace in ancient times of a distinct legal profession in the modern sense. The earliest known legal specialist was the judge, and he was only a part-time specialist. The chief, prince, or king of small societies discharged the judicial function as part of the general role of political leader. As his power spread, he delegated the function, though not to legal specialists; in the secular stages of the early systems, legal duties were taken over by royal officials who were "generalists." In the wake of powerful religious or quasi-religious movements priests or wise men often judged or advised the judges, a situation that persisted in Muslim countries and in China until the 20th century AD. It may be suspected that in some of these cases specialized legal aid to the ordinary citizen did exist, but at levels of social status below the notice of chroniclers or tomb inscriptions and perhaps without benefit of official approval.

Classical beginnings of a legal profession. A distinct class of legal specialists other than judges first emerged in the Greco-Roman civilization, and as with the law itself, the main contribution was from Rome in the period from 200 BC to AD 600. In the early stages of both Greece and Rome, as later among the German tribes who overran the Roman Empire, there was a prejudice against the idea of specialists in law being generally available for fee. The assumption was that the citizen knew the customary law and would apply it in transactions or in litigation personally with advice from kinsmen. As the law became more complex, men prominent in public life—usually patricians—found it necessary to acquire legal knowledge, and some acquired a reputation as experts. Often they also spent periods serving as magistrates and in Rome as priests of the official religion, having special powers in matters of family law. Among the German tribes noble experts were allowed to assist in litigation, not in a partisan fashion but as interpreters (*Vorsprecher*) for those unready of speech who wished to present a case. The peculiar system of development of the early Roman law, by annual edict and by the extension of trial formulas, gave the Roman patrician legal expert an influential position; he became the jurisconsult, the first nonofficial lawyer to be regarded with social approbation, but he owed this partly to the fact that he did not attempt to act as an advocate at trial—a function left to the separate class of orators—and was prohibited from receiving fees.

The jurisconsult and his successors

The modern legal professional, earning his living by fee-paid legal services, first became clearly visible in the later Roman Empire when the fiction that a jurisconsult received only gifts was abandoned and when at the same time the permissible fees were regulated. Changes in the methods of trial and other legal developments caused the jurisconsult to disappear in time. The orator, who now was required to obtain legal training, became the advocate. A subordinate legal agent of the classical system, the procurator, who attended to the formal aspects of litigation, took on added importance because later imperial legal procedure depended largely on written documents drawn by procurators. The jurisconsults had been important as teachers and writers on law; with their decline this function passed to government-conducted law schools at Rome, Constantinople, and Berytus and to their salaried professors. There was also a humbler class of paid legal documentary experts, the *tabelliones,* useful in nonlitigious transactions.

Medieval Europe. This late Roman pattern of legal organization profoundly influenced the Europe that began to arise after the barbarian invasions from AD 1000 on; and even during the invasions the methods of Roman imperial administration never ceased to exist in some parts of southern France and in central Italy. The Christian Church, which became the official Roman imperial church after AD 381, developed its own canon law, courts, and practitioners and followed the general outline of later Roman legal organization. Because of its success among the invaders the church was in a position to establish its jurisdiction in many matters of family law and inheritance. Hence both the idea of a legal profession and the method of its operation retained sufficient force to offset Germanic and feudal objections to legal representation. After the revival of learning in the 12th century, in particular the revived study of Roman law at Bologna, the influence of the late Roman professional system was greatly strengthened.

Classes of legal practitioners

From then on every country in continental Europe acquired, by various stages and with numerous local variations, a legal profession in which four main constituents could be observed. Procurators attended to the formal and especially the documentary steps in litigation. Advocates, who usually were university graduates in Romanist learning, gave direct advice to clients and to procurators and presented oral arguments in court. Among a miscellany of legal scribes the notaries acquired importance because, in addition to being drafting experts, they also provided officially recognized document authentication and archives. University teachers of law took over the main task of explaining and of adapting the mixture of Roman law and Germanic custom that produced the modern laws of the major European countries and continued to dominate in the scholarly interpretation of the law even after the 19th-century codifications. The relative importance of these classes varied enormously from place to place and from century to century. At times the teaching doctors almost supplanted the advocates; in some courts the procurators swallowed up the advocates and in others the converse occurred; only the notaries managed to survive with little change.

England after the Conquest. England after the Norman Conquest also was influenced by Roman example, and the clerics who staffed the Norman and Plantagenet monarchies and who provided the earliest of their judges enabled the notion of a legal profession and especially of litigious representation to be accepted. Only in the ecclesiastical and admiralty courts, however, did procurators (proctors) and doctors of the civil and canon laws become established as practitioners. The native "common law" was developed by a specialized legal society, the Inns of Court in London; there, through lectures and apprentice training, men acquired admission to practice before the royal courts. More particularly, they could become serjeants—the most dignified of the advocates, from whom alone after about 1300 the royal judges were appointed. Various agents for litigation resembling procurators also became known. The "attorneys," authorized by legislation, at first shared the life of the Inns with the "apprentices" in advocacy, who themselves in time acquired the title of barristers. Indeed there were cases of men working as both barristers and attorneys. When in the 16th century the Court of Chancery was established as the dispenser of "equity," the appropriate agent for litigation was called a solicitor, but the common-law serjeants and barristers secured the right of advocacy in that court. It was not until the 17th century that the attorneys and solicitors were expelled from the Inns and the division between advocate and attorney became rigid, and not until the 18th century that the barristers accepted a rule that they would function only on the engagement of an attorney—not directly for the client. Other types of legal agents also developed in England, but in the 19th century all of the nonbarristers were brought under the one name, solicitor. The order of serjeants was wound up, leaving only barristers, of whom the most senior could be made Queen's (or King's) Counsel.

The English division

In its final development the English legal profession thus bore a resemblance to the European—particularly to that of northern France, where the *parlements* (courts) had a corporate life and apprentice training not unlike that of the Inns. But there were four significant differences between England and the Continent. No distinct class of university teachers and commentators on the national law developed in England. Development of the law took place chiefly through precedent based on the reported judgments of the courts, rather than through legislation. The continental monarchies also developed a system of career judicial office, in which the young university licentiate went straight into government service, whereas in England appointment of judges from the senior practicing profession remained the settled practice. In addition, the division between barristers and solicitors ultimately became much more rigid in England than did the division between the advocate and procurator in Europe, and Europe never adopted an equivalent of the English practice requiring a barrister to be employed by a solicitor; both the procurator and the advocate were separately and directly employed by the client. England never developed the profession of notary, so that the whole burden of transactional work fell on those who are now the solicitors, with legal advice from the bar.

Worldwide legal profession. The main patterns both of law and of legal practice were exported by the continental European powers and England to their overseas colonies and possessions, and most of the noncolonial countries of the rest of the world imitated one or the other system. Thus the Romano-Germanic practices (frequently called civil law) became the norm for Scandinavia, Scotland, Latin America, and most of the Muslim countries of the Middle East, for French-speaking areas and Portuguese and Spanish Africa, and for Japan, Thailand, and the former French parts of Southeast Asia. They have also influenced practice in what are now the socialist countries of eastern Europe. The English system provided the model for English-speaking North America, for most former English colonies in Africa, including South Africa, for most of the Indian subcontinent, and for Malaysia, Australia, and New Zealand. The original model has undergone considerable modification by both the countries of export and the countries of reception. In particular, the specialization of procurator–advocate and solicitor–barrister has tended to be replaced by a "fused" profession of legal practitioners qualified to perform both functions and usually doing so. Such a fusion occurred gradually in Germany between the 16th and 18th centuries. It has taken place more recently in France except before the courts of appeal and, while the division still formally exists in Italy, it is no longer of practical importance. In Latin America the fused profession is general. Notaries as a separate specialized branch of the profession exist, however, in most civil-law countries.

CHARACTERISTICS OF THE PROFESSION

Social role. The legal profession has always had an ambiguous social position. Leading lawyers have usually been socially prominent and respected—the sections of the profession so favoured varying with the general structure of the law in the particular community. The family status of early Roman jurisconsults may have been more important than their legal expertise in securing such a position, but by the time of the principate it was their legal eminence that made them respected. The English serjeants lived magnificently, especially in Elizabethan times, and the French Ordre des Avocats was established (14th century AD) by feudal aristocrats in circumstances reminiscent of early Rome—including an insistence on receiving gifts rather than fees. The early Italian doctors of the civil and canon law (12th–15th centuries) were revered throughout Europe. In England and the countries influenced by its system the highest prestige gradually came to be concentrated on the judges rather than on the order of serjeants, of which they were members, and the judges of high-level courts remain the only legal class in the liberal capitalist common-law countries of today to command great respect. In the Romano-Germanic systems it is the notaries and the advocates who have come to be most trusted or admired, the judiciary being more closely identified with the civil service.

The legal profession and social conflict

Yet along with this high repute, sustained over two millennia, lawyers have also been among the most hated and distrusted elements in society. In a few cases this has been the consequence of a general hostility to the whole idea of law, China being the most important example. Confucian teaching (6th century BC) opposed the use of civil law as a major means of social control, and this influence remained powerful there and in Japan until the 20th century. In the Soviet Union the early leaders (1917–22) imagined that law and lawyers were the instruments of the ruling classes and that law would soon wither away in classless Communism. This belief was revived in the early days of Chinese Communism (1947–55). Further experience persuaded these governments that there was room for "socialist legality" and for lawyers to serve it, but a degree of mistrust remains and the repute of the legal expert is lower than that of the political and technological expert.

Most lawyers are conservative because the law itself is predominantly intended to satisfy expectations arising from an inherited pattern of behaviour; in a particular social setting this tends to identify the lawyer with the established and successful classes and to make him seem an enemy to oppressed classes or "new men." Individual lawyers have, nevertheless, often been on the side of rebels: Robespierre and Lenin were both lawyers. But the dominant attitude of the legal profession is one of moderation. Thus many lawyers took the British side in the American Revolution, and even among the lawyers who took the other side the predominant influence was against any attempt to turn the political revolution into a social revolution.

Along with these ideological and political reasons for popular distrust, and even more deep-seated, are the inherent difficulties associated with law and with some of the legal functions. Most people would like law to be so certain that its application is of equal certainty in all cases and so simple that any person of sense can see how it applies. In a discipline sharing the imperfection and complexity of society itself, no such situation is attainable, and the lawyers are blamed for the basic difficulty of their craft—which in some instances they intensify needlessly themselves by multiplying obscurities, contradictions, and complexities. The legal function likely to be most distrusted by the average person, though it also produces some of the law's heroes, is litigious advocacy, particularly in the criminal law. Plato and Aristotle condemned the advocate as one who was paid to make the worse cause appear the better or endeavoured by sophisticated tricks of argument to establish as true what any person of common sense could see was false. The feeling against advocacy in the criminal law was so strong that, at least in the case of the more serious kinds of crime, a right to representation by a trained advocate was nowhere generally recognized until the 18th century AD.

Governments and the members of organized legal professions have from the infancy of the craft endeavoured to meet the basic problem of representation by a basic rule of professional ethics—that the dominant duty of the advocate is not to his client but to truth and the law. Since the later Roman Empire, advocates have been required to take oaths to this effect, and lawyers are often technically classed as "officers of court." The duty of the advocate is to fight for the rights of his client, but only up to the point where an honourable person could fairly put the case on his own behalf. He must not identify with his client's possible willingness to tell untruths or to misrepresent the relevant law. (See below *Legal ethics.*)

Private practice. Client-directed lawyers often are called counselors, but in the original sense of that word—giving advice as to how the law stands—this is rarely an independent function; it is an inseparable part of the other functions. In his client-directed activities the lawyer is concerned with how the law affects specific circumstances, which can for convenience be divided into two main types: transactional and litigious.

In the transactional type the lawyer is concerned with the validity or legal efficacy of a transaction independent of any immediate concern with the outcome of litigation. Such activities comprise the largest area of professional activity whether considered from the point of view of the number of lawyers involved, or of the time they have to spend on the task, or of the number of clients affected. If the events constituting the transaction in question happen before the lawyer is consulted, he can only advise on their legal significance and perhaps suggest methods of overcoming legal deficiencies in what has been done. If future conduct is involved, he is better placed to plan his client's course of conduct so as to secure the required end in the most economical fashion that the law permits and to minimize the chances of future litigation. Transactions may concern words and acts, but characteristically they require the drafting of documents. In the Romano-Germanic systems these often require notarization. Typical activities falling in this category today include the following: transferring interests in land; transmitting property on death; settling property within a family; making an agreement (especially if a commercial agreement of some complexity and duration is involved); incorporating or winding up a corporate entity; varying the terms on which a corporate entity is conducted (classes of shares, managerial rights, distribution of profits, etc.); and adjusting the ownership and control of property and income so as to comply with the requirements of taxation laws and minimize their impact on the property and income in question or so as to ensure the proper management of the assets and distribution of the proceeds among beneficiaries (estate planning) or both. In the Romano-Germanic systems many of these functions are discharged by notaries and in the English and similar divided systems by solicitors, though in difficult situations the opinions of advocates or barristers may be obtained. In the fused professions of North America some firms of attorneys, or departments within firms, specialize in business of this type and avoid, so far as they can, the litigious function.

Typical legal transactions

The litigious function is subdivided into three main stages. First is the preparation of the case—interviewing the client and investigating the circumstances in the light of the leads provided by the client, and attending to the formal requirements of the procedure in question, which may involve writs, summonses, filing of statements of claim or defense, and preparing for trial. Second is the trial proper, in which the facts and law are established and argued before the judge and a decision is made. Third is the execution of the judgment—obtaining payment of damages, delivery of property, or performance of obligation in civil cases; payment of fine or imprisonment, etc., in criminal matters. Similar stages arise on appeal. In the divided professions the sharing of these functions is intricate and varies between one system and another. The advocate or barrister is especially responsible for the second stage, but he may advise upon or draft many of the documents used in other stages. If incidental disputes concerning procedure have to be litigated, he is likely to conduct the proceedings; and, if the procedure includes a pretrial conference, he is likely to represent the client. Otherwise the first and third stages are mainly the province of the procurator or solicitor.

Public-directed practice. Many law graduates choose to enter public service rather than private practice. Of the public roles played by members of the legal profession, that of judge is most visible, but the status of judge and the mode of entry into this branch of the profession vary considerably from country to country.

Status of judges

The traditional independence, power, creativity, and prestige of the Anglo-American judge contrasts with the rather ordinary civil servant status of most continental judges. In the countries of Anglo-American influence, at least until recently, appointment (or, in some U.S. states, election) to a judgeship has been viewed as the crown of a long and distinguished legal career. In the continental countries, by contrast, a law graduate who wishes to be a judge has merely to complete a training period and pass an examination to get a job deciding cases. The beginning civil-law judge can expect to start at the lowest level and, like any other civil servant, to rise in the hierarchy through a series of promotions. In continental Europe ordinarily only positions on the highest courts are open to distinguished practitioners or professors as well as to career civil servants. Lateral entry into the judiciary at any level is uncommon. It has frequently been observed that, because of their standardized training, continental judges tend to share a common outlook and that their concerns about advancement promote a civil service mentality that discourages initiative and independence. Any tendency toward individualism is apt to be further inhibited by the fact that continental judges, even at the lowest levels, usually sit in panels and that their decisions are presented in unsigned opinions. Except in a few courts, such as the German Federal Constitutional Court, disagreement among the judges is not revealed, either in the form of a dissenting opinion or in a record of the judges' votes.

In the late 20th century, however, the contrast between continental and Anglo-American judicial roles has diminished. In the United States the prestige of judgeships, except at the higher levels, has declined somewhat. It is not unusual for judges to resign and return to private practice or for eminent lawyers to decline to be considered for judicial positions. Relatively low judicial salaries are often mentioned as a reason why lower court judgeships in the United States have ceased to attract the best candidates. Meanwhile, in some continental countries, such as Germany, judges are being recruited from among the best law graduates. Because of the special training for their positions, continental judges are almost uniformly professional and competent.

Governments have always required a staff of legal specialists, and the scope for such employment today is enormous. There are usually a senior political officer—minister of justice, attorney general, solicitor general—who by convention needs to be a lawyer, and a government department concerned mainly with the legal problems of the government as client (in the English-derived systems usually the office of the attorney general). Increasingly, however, the great departments of state need their own legal subbranch. In some countries, Germany for example, lawyers dominate the higher offices in the civil service, while in others, such as France, the various official bureaus are more likely to be staffed by nonlawyers who have been trained in a special school of administration. In the formerly socialist countries of eastern Europe, most lawyers tended to work for government or for collectivized industrial and farm organizations.

One of the oldest and still most difficult of the governmental legal functions is that of prosecutor in criminal cases. Prosecution is sometimes in part carried on by private persons acting through private lawyers, but the trend is to concentrate the function in government legal officers. In most Commonwealth countries the crown, or public, prosecutor is a specialized officer under the general control of the attorney general. England has an independent "director of public prosecutions" concerned only with the most serious types of crime, but most prosecutions have been conducted by private barristers briefed by him or by the police. A 1985 law, however, provides for the establishment of a body of official prosecutors, following the Scottish system, which relies on public prosecutors (procurators fiscal). In the United States this function has come to be mainly local, and prosecutors, whose most common title is district attorney, are elected for short terms.

In the Romano-Germanic systems prosecuting is a career service. In Italy and France the prosecutor is a member of the judiciary. Both prosecutors and judges receive the same training, and both may move from one role to the other in the course of their advancement in the civil service. In Germany, although the prosecutor is not technically a member of the judiciary, he is not strictly separate from it, and individuals move easily from one position to the other.

The prosecuting function is particularly delicate because criminal prosecution can be used as an instrument of oppression and political persecution, even where conviction is not obtained, and because in most systems prosecutors are expected to act with a degree of fairness and restraint not necessarily expected of the parties to civil litigation. Many Romano-Germanic systems employ officers who keep a general supervision over the working of the courts and especially their criminal jurisdiction. This is the office of the "prosecutor general," or "officer of justice"; a similar service existed in most of the socialist countries of eastern Europe.

Another branch of government, the legislature, usually requires legal assistance. Legislation needs to be expressed in language readily comprehensible by judges and lawyers and to be framed in harmony with the existing body of law. This requires the service of parliamentary draftsmen who are expert lawyers. A further specialized branch of advisory activity associated with legislation has become prominent—the law reform commission or committee.

Teaching and scholarship. Since Roman times teaching and scholarship in the law have provided prominent roles in the legal profession. Until the 18th century, teaching of the English common law was vested exclusively in the Inns of Court, and a good deal of continental European teaching for professional practice—particularly in the case of notaries and procurators—was also professionally organized. Even university law teaching in Europe was rarely aloof from legal practice; there was usually a fruitful interchange between practitioner and teacher, exemplified in such great figures as the French 18th-century teacher, advocate, and judge Robert Joseph Pothier, whose commentaries provided the foundation for the Napoleonic Code of civil law.

University law schools

Much law teaching in the new university law schools that sprang up in the United States, the United Kingdom, and the Commonwealth in the 19th and 20th centuries was carried on part-time by attorneys, barristers, and judges, and some still is. Sir William Blackstone, the first Vinerian professor of English law at Oxford, came from the bar and became a judge. Law teaching in the late 20th century has tended to become a distinct, full-time profession, usually carried on at a university.

Teachers and practitioners in all countries contribute to an enormous professional literature, with students' texts, practical manuals, theoretical monographs, and a periodical literature whose bulk is coming to be almost as big a problem as the enormous bulk of reported judicial decisions that are consulted for guidance and precedent. Civil-law judges pay close attention to the views of legal scholars as expressed in general and specialized treatises, commentaries on the codes, monographs, law review articles and case notes, and expert opinions rendered in connection with litigation. Persistent scholarly criticism often prompts reexamination of a legal doctrine and sometimes even leads to the abandonment of an established judicial position. In the Anglo-American systems certain kinds of legal writing, such as leading treatises, have become highly influential, as evidenced by the measurable increase in citations to secondary sources in contemporary judicial opinions. But the degree of deference to academic opinion is in general much less than in the continental countries.

Trends in the profession. Beginning in the 1960s, nearly everywhere, the legal profession experienced unprecedented expansion, which began to slow down in the 1980s. Major elements in this expansion were the postwar baby boom, the opening of new university law faculties, and the entry of women into the profession in significant numbers for the first time. The growth of the profession has increased competitive pressures for jobs and business.

Trends toward specialization and bureaucratization are also clearly noticeable in the modern legal profession. Although formal divisions among lawyers based on functions have declined, there tends to be in most places a de facto division of labour between those lawyers who advise clients and those who appear and argue before tribunals as well as increasing specialization in various legal fields, such as tax law, estate planning, and labour law. Bureaucratization has affected both the practice of law and the judiciary. In all countries the ranks of those lawyers who work as salaried employees of government, business, or law firms are growing more rapidly than those of traditional independent professionals. Changes in the administration of justice in the United States, brought about by increased litigation and crowded dockets, have bureaucratized the roles of judges. Courts in the United States increasingly depend on large and layered staffs to perform many functions that were previously considered to be within the province of the independent judge. While judges in the United States are far from being civil servants in the continental European sense, their work is becoming more administrative in nature. (G.S./M.A.Gl.)

AUTONOMY AND CONTROL

Issues of judicial independence. At least since classical Greece a recurring political theme has been the need for a government of laws rather than of men. Actually, however, as the legal philosopher Julius Stone has said, society of necessity has a government both of laws and of men, and the demand for legal autonomy is often seen in practice as a demand for freedom of the lawyers from political dictation or influence. The main issue has been the independence of the judiciary, and democracies have been particularly assiduous in cultivating both a spirit and traditions that respect judicial independence. The details of their governmental structure or constitutional guarantees tend in that direction, offering obstacles to the ready dismissal of judges, charging their salaries on consolidated revenue, and prohibiting the vesting of judicial functions other than in duly constituted courts of law.

Political context of the judiciary

The special position of the judiciary in constitutional states is usually considered to be an aspect of the division of powers, but it also should be considered in its relation to the structure of the legal profession. Since the late Roman Empire, admission to the practice of law and the regulation of the practicing profession have been habitually vested in the judiciary. Furthermore, the duty to speak fearlessly for his client has often required courage of the advocate in the face of political threats directed against him, and, when these threats were directed also against the court before which he appeared, judicial courage also was required. The legal profession as a whole is then seen defending "the rule of law" against the political regime.

The issue of judicial independence may sometimes, however, be seen in the context of tension between the judges and the advocates. In the Romano-Germanic systems the judges often are subject to a strong corporate discipline within their own craft, and differences can occur between them and the body of advocates and also between them and the university teacher-commentators. These differences may relate to questions of legal ethics, especially the limits of advocate identification with client, or to questions of legal doctrine; the judges are then apt to be considered as representing "the state," and the advocates and teachers the autonomy of the law. In the English-derived systems judges are much less subject to corporate discipline, and disputes with the bar are more likely to arise with individual judges and to be highly personal. Even in stable countries, where the rule of law and the independence of judiciary and profession are respected, there is a less dramatic tension between the standards and tone of the lawyers on the one hand and the political administration on the other. For the lawyers, policy is largely concealed in the propositions that constitute the normative system, and legal reasoning usually involves definitions and processes of inference from the body of such propositions themselves rather than directly from the policies that the norms subserve. There have often been revolts against such "logic" within the legal profession itself, especially in the 20th century, but it still remains the most common method of thinking among lawyers, and it is doubtful whether one can speak of a "rule of law" at all unless a good deal of legal reasoning is conceptual in style. Politicians and administrators, on the other hand, are more likely to reason directly from policies and purposes and from the considerations relevant to their attainment. This divergence of approach is often illustrated by referring to the tension beween the police officer, confident that he has the guilty man and intent only on putting him in jail, and the lawyers and judge, who insist on the need for "conviction according to law," which may involve applying rules of evidence that seem artificial and even absurd to the police officer. In rigid constitutional systems, where there is judicial review of legislation, politicians may be affronted at the way in which political issues are transformed by the lawyers into legal issues. In many modern countries there has been a tendency to remove certain kinds of disputes both from the courts and from the lawyers and to vest their determination in administrative bodies before which lawyers are denied standing, so as to escape what has been regarded as the blight of legal reasoning; as often there have been reactions in favour of restoring the "rule of law" and the lawyers. In such disputes it is often difficult to distinguish between lawyer attitudes that reflect the necessary features of a rule of law from those that merely reflect the temporary self-interest of particular lawyers or their clients.

Regulation by statutes and bar associations. Since about 1800 most countries have brought their legal professions under systems of statutory control with three main principles: admission to practice automatically and compulsorily makes the lawyer a member of an appropriate professional association; those associations are given substantial powers in relation to legal education, admission to practice, and the disciplining of the profession, but they are subject to overriding powers vested in the courts and/or (especially in the Romano-Germanic systems) in government legal departments; the practice of law for reward is prohibited—generally or as to particular functions—to persons not admitted under the system. In the United States about half of the states have such a system, which is known as the "integrated bar"; in the other states bar associations are voluntary and have few controlling powers. England has retained the traditional Inns of Court (in whose management the judges play a leading role) for barristers, but solicitors are subject to a statutory system as above. In some countries (*e.g.*, France) professional organization is regionalized to correspond with judicial organization, and in some federal countries (*e.g.*, the United States, Canada, and Australia) professional control is vested in the states; such situations create the problem of a national organization that is generally a voluntary federation of regional bodies and therefore lacking in compulsive authority. The American Bar Association, established in 1878, is a leading example. In other federal countries (*e.g.*, Germany and India) the central government has created national law associations responding to the need for a system of control. The law associations, apart from the functions already mentioned, help their members to understand and apply professional ethics, and they develop canons of ethics to cover new problems. They are often active in the prohibition of legal practice by unqualified persons, which tends to bring them into dispute with other professions—*e.g.*, tax accountants and land salesmen—whose members wish to perform legal functions in relation to their tasks and often have considerable knowledge of the relevant law.

Professional associations: protection and control

Where the profession is divided, it is usually possible to transfer from one branch to another, though sometimes after delay or additional training. In many of the Romano-Germanic systems, however, professional mobility is severely restricted by another factor—numerical limits on the numbers admitted to a branch of the profession.

There are usually limits to the numbers of procurators and notaries, and in some cases, notably the highest French courts, advocate and procurator functions have been combined in relation to a particular jurisdiction and a limit has been placed on numbers; otherwise the number of advocates is generally not restricted. In the restricted cases a person admitted to practice can actually work in the profession only as an employee of an existing practitioner or after buying out such a practitioner.

The actual ability to enter or pursue the profession can also be much influenced by the varying national rules as to legal partnerships. They are prohibited for English barristers and for most divided bars derived from that system and among some of the Romano-Germanic specialized advocates and notaries. In France law partnerships are permitted, and the proportion of lawyers practicing in this manner is constantly increasing. Incorporation of legal practitioners is almost universally prohibited. These restrictions result from the emphasis on personal responsibility of the individual lawyer to his client, to the court, and to the ethical system. In countries with fused professions, however, partnership is usually permitted. German law firms tend to be quite small, and even in the United States, despite the fact that some very large firms have developed, particularly among the "corporation lawyers" of New York, single-person practices and small partnerships are still common. (G.S./M.A.Gl.)

Legal ethics

Sources of legal ethics

Legal ethics are the principles of conduct that members of the profession are expected to observe in the practice of law. They are an outgrowth of the development of the legal profession itself.

Practitioners of law emerged when legal systems became too complex for those affected by them to understand and apply the law. Certain individuals with the required ability mastered the law and offered their skills for hire. No prescribed qualifications existed, and these specialists were not subject to legal controls. The incompetent, unscrupulous, and dishonest charged exorbitant fees, failed to perform as promised, and engaged in delaying and obstructive tactics in the tribunals before which they appeared. Action to prevent such abuses was taken by legislation and by judicial and other governmental measures. The right to practice law came to be limited to those who met prescribed qualifications. Expulsion from practice and criminal penalties were introduced for various types of misconduct.

These measures did more than correct the abuses. They also gave recognition to the social importance of the functions performed by lawyers and identified those who were qualified to perform them. A consciousness developed within the profession of the need for standards of conduct. This became the core of legal, or professional, ethics.

Prior statutes, court rules, and other government directives remained in force along with the profession's self-imposed ethical standards. Taken together, they constituted the sum total of the restraints placed upon lawyers in regard to their professional conduct. This pattern has continued to the present time.

In many countries professional associations of lawyers have sought to commit the principles of ethical conduct to written form, but a written code is not essential. Ethical principles may exist by common understanding as well as in the literature and writings of the profession. This is the case in England. A code, however, makes ethically obligatory principles readily available to the practitioner and thus helps to assure greater observance of them. When such a code does exist, it usually contains both statements of general ethical principles and particular rules governing specific problems of professional ethics. But no code can foresee every ethical problem that may arise in the practice of law. Hence, codes are supplemented by opinions rendered and published by committees of bar associations.

DIFFERENCES AMONG COUNTRIES

Principles of legal ethics, whether written or unwritten, not only seek to control the conduct of legal practice but also reflect the basic assumptions, premises, and methods of the legal system within which the lawyer operates. They reflect as well the profession's conception of its own role in the administration of justice. In western European legal systems the role of the practitioner in both civil and criminal litigation differs from what it is in Anglo-American legal systems, and this is reflected in the ethics of their legal professions. Practitioners in countries having a Communist form of government are ordinarily salaried employees of the government, and their ethical obligations consequently have a focus different from that in countries where lawyers engage in independent practice or are employed by private firms. In England and to a certain extent in France, where the profession is divided into separate branches, the principles of professional ethics reflect the relationships incident to that division.

In western European and Anglo-American countries and others with similar systems of justice, such as Japan and India, in which the lawyer is not an employee of the state but engages in private practice to serve clients who employ him, professional ethics are addressed to two basic aspects of the lawyer's status. On the one hand, he is employed by clients to serve and represent their interests; on the other, he is participating in an important social function—the application of rules of law through advice, trial of cases, preparation of legal documents, and negotiation with others for his clients. Hence, the principles of legal ethics stress that the lawyer's chief interest lies in serving his client and in securing justice—not in increasing his own income. He is an agent of his client but deemed to retain a large measure of independent judgment as to the proper course to pursue. He represents his client's interests but may not engage in tactics that defeat the fair administration of justice. The lawyer is engaged, it is said, in a profession and not in a business.

Conflicting interests of client and society

Naturally the interests of client and society sometimes conflict, and the principles of legal ethics do not always indicate how these conflicts should be resolved. Should a lawyer cross-examine an adverse witness in such a way as to undermine or destroy his testimony when the lawyer believes the witness is actually telling the truth? May he invoke rules of evidence to exclude points that would weigh against his case but that he considers to be probably true? May he take advantage of the errors of an unskilled opponent? Should he demand a jury trial for purposes of delay when a jury trial has no advantage for his client? These questions may be answered differently in legal systems that operate on different premises. A system in which a lawyer presents his client's case in the most favourable light permitted by law and in which the court must decide the merits of the case may well produce different answers than those produced in a system that assigns a higher priority to the lawyer's duty to the state to assure proper administration of justice.

AREAS OF APPLICATION

Conflicting interests. A lawyer is at times faced with the question of whether he may represent two or more clients whose interests conflict. Quite aside from his ethical obligation, the legal systems of the world generally forbid a lawyer from representing a client whose interests conflict with those of another, unless both consent.

In the Anglo-American legal systems the prohibition has three aspects. First, the attorney is not permitted to concurrently represent two or more clients if, in order to further the interests of one, he must forego advancing the conflicting interests of another. In short, he cannot be both for and against a client. Second, he cannot subsequently accept employment from another for the purpose of undoing what he had earlier been retained to accomplish. Third, he may not accept subsequent employment from another if it involves the use, the appearance of use, or possible use of confidential information received from his former client. Such actions are forbidden by law and by legal ethics.

To illustrate, an attorney may not ordinarily prepare an instrument for both buyer and seller in which their respective rights are defined. He may not prepare an instrument or negotiate a settlement for a client and later accept employment from another to defeat that instrument or settlement. He may not represent both a driver and his passenger in recovering damages from another party charged with negligent driving in a collision since the passenger may have a claim against his own driver as well. He may not represent two or more defendants in a criminal prosecution if their respective defenses are inconsistent or, possibly, even when the case against one is stronger than the case against the other. The same principles apply with respect to interests of the attorney that may detract from the full and faithful representation of his clients. For example, he may not purchase property that he has been retained to acquire for his client, nor may he draw a will in which he is a beneficiary.

These conflict-of-interest prohibitions are not absolute. The client may consent to the representation after full disclosure of the actual or possible conflict. But the client's consent may not suffice if public interest is deemed to be adversely affected.

The practicing lawyer who is also a member of a legislature is confronted with a conflict of interest whenever his clients enlist his support to promote or oppose legislation or to secure favourable decisions from administrative agencies that are dependent on legislative financial support. The problem is an important one in the United States, where members of legislatures frequently maintain private law practices, but it has received insufficient consideration by the U.S. legal profession. (M.E.P.)

Confidential communications. In the Anglo-American countries judicial decisions, legislation, and professional ethics forbid a lawyer to testify about confidential communications between himself and his client unless the client consents. Similar provisions are found in such diverse legal systems as those of Japan, Germany, and Russia. In countries in which the attorney's obligation to protect state interests is given relatively greater emphasis, there may be a duty to disclose information when it is deemed to be to the state's advantage. In Anglo-American law the obligation does not apply when the client seems about to commit a crime. An attorney also may disclose his client's communication when the client sues him—for example, for malpractice.

Advertising and solicitation. Traditionally, advertising by lawyers was forbidden almost everywhere. It has been a long-standing principle of professional ethics in Anglo-American countries that an attorney must not seek professional employment through advertising or solicitation, direct or indirect. The reasons commonly given have been that seeking employment through these means lowers the tone of the profession, that it leads to extravagant claims by attorneys and to unrealistic expectations on the part of clients, and that it is inconsistent with the personal relationship that should exist between attorney and client. A more basic reason appears to have been the social necessity of restraining the motive of personal gain and of stressing the objective of service. Until 1977 the legal profession in all Anglo-American countries took the position that, with some exceptions, the prohibition must be complete. The situation changed in the United States in 1977 when the U.S. Supreme Court ruled that lawyers could not be barred from advertising fees. The American Bar Association subsequently revised its code of ethics to include provisions and guidelines for advertising and suggested that lawyers limit their advertising to basic information about services and fees. Within narrow limits the same trend has made itself felt in England.

Dignity of the profession

Fees. Attorneys are ethically enjoined to keep fees reasonable, neither too high nor too low. Attempts to control fees range from mandatory fees fixed by statute in Germany, minutely regulating compensation for legal services of all sorts, to mandatory fees set by courts for solicitors in contentious matters in England and Wales, to advisory fee schedules established by the profession in Canada, France, Spain, and Japan. In the United States local bar associations sometimes enforced minimum fee schedules through disciplinary proceedings; however, the U.S. Supreme Court held in 1975 that such practices violated the antitrust laws.

The profession in the United States has assumed, in principle, the obligation to serve poor clients without compensation. The task, however, has become so enormous, especially in view of the expansion of the constitutional right to counsel in criminal cases, that ways of providing paid legal services for the poor have emerged, such as through legal aid societies and public defenders. The growth of legal aid has been a significant 20th-century development in many other countries. In Germany legal insurance plans are widespread as well.

Fees that are contingent on the successful outcome of litigation or settlement are widely used in the United States, particularly in automobile-accident and other negligence cases, and they are accepted as ethical by the U.S. legal profession. The fee is usually an agreed percentage (typically 20 to 40 percent) of the recovery. The justification given is that this arrangement makes the courts accessible to persons who would otherwise be unable for financial reasons to press their claims. But contingent fees give the attorney a financial stake in the outcome of litigation—which is ordinarily frowned upon. The converse consideration may be that in this type of case, where the outcome is difficult to predict, the lawyer also assumes the risk of losing his fee. Furthermore, although free legal aid has removed the need for a poor person to enter into such a transaction, legal aid is not available to persons who are not poor but are not wealthy enough to engage in extended litigation. In countries other than the United States contingent fees are, nevertheless, generally prohibited. Nor are they permitted in the United States in criminal and divorce cases, in cases to secure a pardon, or in the enactment of legislation.

Honesty in court

Criminal cases. Both the prosecution and the defense of criminal cases raise special ethical issues. The prosecutor represents the state, and the state's concern is not only in convicting the guilty but also in acquitting the innocent. The prosecutor also has an ethical and, in considerable measure, a legal duty to disclose to the defense any information known to him and unknown to the defense that might exonerate the defendant or mitigate the punishment. He must not employ trial tactics that may lead to unfair convictions, nor should he prosecute merely to enhance his political prospects.

The defense counsel has different concerns. Under Anglo-American law an accused may compel the state to prove that he is guilty beyond a reasonable doubt. The defense counsel, therefore, becomes ethically obligated to require the state to produce such proof, whether or not the attorney believes his client to be guilty. His client's guilt is for the tribunal to determine. The attorney may not, however, deliberately resort to perjured or other false testimony. Similar principles hold in civil-law countries. When the client, against the attorney's advice, insists on testifying falsely, the ethical course to be pursued has not been fully settled. Some maintain that the attorney should withdraw, if possible, or else merely permit the client to testify without aiding him or asserting the truth of the testimony given. (M.E.P./M.A.Gl.)

Legal education

The rise of legal education

Schools of law are of comparatively recent origin. The ancient Romans had schools of rhetoric that provided training useful to someone planning a career as an advocate, but there was no systematic study of the law. During the 3rd century BC, Tiberius Coruncanius, the first plebeian *pontifex maximus* (chief of the priestly officials), gave public legal instruction, and a class of nonpriests (*jurisprudentes*) who acted as legal consultants emerged. A student, in addition to reading the few lawbooks that were available, might attach himself to a particular *jurisprudens* and learn the law by attending consultations and by discussing points with his master. Over the ensuing centuries a body of legal literature developed, and some *jurisprudentes* set themselves up as regular law teachers.

In the medieval universities of Europe, including England, it was possible to study canon law and Roman law but not the local or customary legal system. The study of national laws at universities is in most European countries a development that began in the 18th century; the study of Swedish law at Uppsala dates from the early 17th century.

On the continent of Europe the transition to the study of national law was facilitated by the fact that modern legal systems grew mostly from Roman law. In England, on the other hand, the national law, known as the common law, was indigenous. In medieval times education in the common law was provided for legal practitioners by the Inns of Court through reading and practical exercises. These methods fell into a decline in the late 16th century, mainly because students came to rely on printed books, and after the middle of the 17th century there was virtually no organized education in English law until the introduction of apprenticeship for solicitors in 1729. The famous jurist Sir William Blackstone lectured on English law at Oxford in the 1750s, but university teaching of the common law did not develop significantly until the 19th century. The Council of Legal Education for barristers was established in 1852. In the United States systematic legal education began with the founding of the Harvard Law School in 1817. (L.A.S.)

THE AIMS OF LEGAL EDUCATION

Legal education generally has a number of theoretical and practical aims, not all of which are pursued simultaneously. The emphasis placed on various objectives differs from period to period, place to place, and even from one teacher to another. One aim is to make the student familiar with legal concepts and institutions and with characteristic modes of reasoning. Like most intellectual disciplines the law has its technical concepts, frequently expressed in technical terms. All lawyers must become acquainted too with the processes of making law, settling disputes, and regulating the legal profession. They must study the structure of government and the organization of courts of law, including the system of appeals and other adjudicating bodies.

Another aim of legal education is the teaching of law in its social, economic, political, and scientific contexts. While law schools have never ignored the social context of their subject, Anglo-American legal education has always been less interdisciplinary than that of continental Europe. With the development of a more or less scientific approach to social studies in the 20th century, however, this is changing. Some law schools appoint economists, psychologists, or sociologists to their staffs, while others require or permit their students to take courses outside the law school as part of their work toward a degree. This awareness of the other social studies is thought to be more advanced in the United States than in Great Britain. Continental legal education (in both eastern and western Europe) tends to be highly interdisciplinary, with nonlegal subjects compulsory for students taking their first degree in law.

Legal history

Traditionally, legal education included the teaching of legal history, which was once regarded as an essential part of any educated lawyer's equipment. While legal history has lost prestige in the sense that separate courses in the subject are offered in few law schools and, when optional, are not very popular among students, much legal history is, nonetheless, taught in the context of other courses. Since the corpus of the law is a constantly evolving collection of rules and principles, many teachers consider it necessary to trace the development of the branch of law they are discussing. In countries where most parts of the law are codified (as, for example, in continental Europe, Central and South America, the countries in the Mediterranean basin and in Africa that were formerly under French influence, Thailand, and Japan), it is not generally thought necessary to go back beyond the codes. On the other hand, in countries that have a common-law system (England, most members of the British Commonwealth, and most parts of the United States), in which few branches of law are codified, knowledge of the law has traditionally depended to a great extent on the study of court decisions and statutes out of which common law evolved. This made the study of legal history of more immediate significance in such countries. However, as the former case-law areas have increasingly come under statutory and administrative regulation, the practical importance of legal history has receded.

The graduating law student is not expected to have studied the whole body of substantive law. He is, however, expected to be familiar with the general principles of the main branches of law. To this end certain subjects are regarded as basic: constitutional law, governing the major organs of state; the law of contract, governing obligations entered into by agreement; the law of tort (or delict), governing compensation for personal injury and damage to property, income, or reputation; the law of real (or immovable) property, governing transactions with land; and criminal (or penal) law, governing punishment, deterrence, rehabilitation, and prevention of offenses against the public order. The chief materials are the same everywhere: codes (where these exist), reports of court decisions, legislation, government and other public reports, institutional books (in civil-law countries), textbooks, and articles in learned periodicals. The aim is not so much that the student should remember "the law" as that he should understand basic concepts and methods and become sufficiently familiar with a law library to carry out the necessary research on any legal problem that may come his way.

STUDY AND PRACTICE

To some extent all law courses are out of harmony with legal practice, for in real life a case is not presented as neatly by a client to his lawyer as it is in a textbook. The case usually begins as a statement, often jumbled, of facts and problems that cut across pedagogical categories. A story of a road accident, for example, may involve the lawyer in considering questions of the civil responsibility for the cause of the accident; of contract (in relation to insurance); of criminal law (in relation to a traffic offense); and of other branches of law as well. It is therefore important, while making divisions of law for convenience of study and examination, to guard students against the danger of thinking in compartments.

Lawyers also must contend in practice with branches of law in which they have received no formal education. More importantly, new social problems requiring legal attention and new legal structures come into existence during every lawyer's lifetime. His task may be eased if he has learned to look to the experience of other nations. A good law school produces a graduate who is not constricted by pedagogy but is trained to adapt himself to—and perhaps lead in bringing about—legal changes related to social, economic, and political developments.

Need for adaptability

The curriculum of the law school also must allow for the great diversity of careers followed by those who have been trained in the law. In most countries large numbers of persons with a legal training seek a career outside the practicing legal profession, commonly in the civil service, in municipal government service, in legal education, and in commerce and industry. Students' requirements and tastes differ; most law schools, therefore, offer a choice. It is common to prescribe a certain number of compulsory subjects, which are regarded as essential to any law student's education, and leave a freedom of selection as to other subjects, stipulating only the number of courses to be studied. With few exceptions, there is little uniformity from law school to law school within the same country as to which subjects are compulsory, and lists of optional subjects vary markedly.

The extent to which legal education aims to teach practice and procedure varies from place to place. Attention is always given to the methods of ascertaining the law from the books but not always to the ways of using this knowledge of the law in various roles, such as legal adviser or judge. Discussion of these matters tends to be more widespread in universities in the United States and in countries where the main qualification to practice the law is a university degree than it is in England and continental Europe, where professional training is provided outside the university and after graduation. In recent years clinical programs, in which students can have real or simulated experience in law practice, have become a staple part of the American law school curriculum. On the Continent such training is typically part of a postgraduate apprenticeship program (see below).

Courses on the rules and principles of court procedure are typically compulsory in university law schools. In England, however, few universities teach these, leaving them to the bar and solicitor's examinations, though the law of evidence (governing what facts may be proved in court, and how) is usually an optional subject; some knowledge of civil and criminal procedure may be picked up incidentally during the study of substantive law.

TEACHING METHODS

Methods of legal education are constantly changing, but the requirement of a university degree has become more or less uniform, coupled in many countries with the need to pass a qualifying examination organized by the profession. Apprenticeship, once a usual way of entering the profession in the common-law countries, has everywhere been increasingly displaced by university education, to which it has now become a supplement.

University law schools tend to differ along national lines in their methods of teaching. In the United States, following pioneer work by Christopher Columbus Langdell at Harvard in the latter half of the 19th century, the case method came to prevail, in which the student reads reported cases and other materials collected in a casebook and the class answers questions about them instead of listening to a lecture by the teacher. The casebook method has been adopted at some institutions in England and other common-law countries but has found scarcely any adherents elsewhere. Even in the United States most law schools now use seminars and lectures as well. The case method has the advantage of emphasizing the characteristic feature of the common law—the evolution of principles from decisions in actual cases—and thus of focusing the student's attention on the processes of analogy and distinction. It has the disadvantages, first, of being relatively time-consuming in relation to the amount of knowledge of legal principle that can be imparted and, second, of concentrating on a source of law that has become just one of many in modern statutory and regulatory legal systems. The traditional teaching techniques in English universities have been lectures and tutorials (or seminars).

Emphasis on the formal lecture

In continental European countries the backbone of legal education is the formal lecture. Class sizes are typically very large compared with those in the United States and Britain, and lectures tend to be magisterial performances. Attendance is frequently voluntary, and those who stay away are usually able to secure the text of what they have missed. Seminars are given too, particularly for specialized subjects. Similar methods are used in other countries with large numbers of law students. In Russia, as in western Europe, the lecture method supplemented by smaller discussion groups is typical. But in Russia the lectures are well attended, and participation in seminars is mandatory.

Teaching methods are not unrelated to the nature of the legal system. The methodology of continental legal education has grown out of and perpetuates a legal tradition heavily influenced by scholars, while the methods in England and the United States have emerged from and contribute to the maintenance of the tradition of judge-made law. Methods were influenced also by the fact that in England legal education was from early times in the hands of the bar, while on the Continent from the 12th century on it was the province of the universities. The fact that in common-law systems principles of law are largely derived by a process of inductive reasoning from many decisions of higher courts lay behind the development of the case method. In continental Europe the fact that law is found mainly in systematic legislation is one of the chief reasons for the lecture method, in which the subject can be approached through its philosophical background. A desire to systematically expound a body of principles rather than approaching facts and problems by the case-by-case method is met better by formal lectures and textbooks than by class discussion. This formal approach is reinforced in countries where published reports of local court decisions are scanty.

Much has happened in recent years, however, to diminish these contrasts. Law schools everywhere are seeking a better balance between theory and practice. American law professors increasingly consider the case method only one of several useful teaching techniques, while continental law faculties, particularly since the student unrest of the 1960s, have instituted various reforms designed to provide students with more opportunities for practical exercises and classroom discussion. The main obstacle faced by continental law faculties engaged in these efforts, however, is the unfavourable ratio of teachers to students.

EXAMINATIONS AND QUALIFICATIONS

The process of selecting members of the legal profession begins in the universities and law schools and continues afterward in the form of professional entrance requirements.

School examinations. In the United States, Great Britain, and other common-law countries students are generally required to pass an examination in each subject. Four or five subjects are studied simultaneously during the academic term, and students must take examinations in all of them at the end of the term or year. In France and Italy, too, students are required to pass a certain number of examinations in various subject matter areas in order to qualify for a degree.

In some continental European countries more comprehensive examinations are the rule. In Germany the course work for the university law degree normally takes about five years, with a single comprehensive examination at the end of those five years (the First State Examination). Students are admitted to this examination if they produce certificates of satisfactory work in each subject, in a jurisprudence seminar, and in a course on economics and finance. The Netherlands has an intermediate system: the course for a first degree in law lasts five years, with an examination at the end of the second year and another at the end of the fifth. Russia combines the system of examinations in each course with a comprehensive examination at the end of the five-year period of study.

Subject-by-subject examination

The method of subject-by-subject examination is less taxing on the memory than the system of comprehensive examination. It may well enable the student to do more detailed work on the problems of each subject. It has the disadvantage of encouraging him to think in terms of separate subjects, whereas the comprehensive examination leads him to consider legal problems in all their aspects. Being aware of the dangers of compartmentalized thinking, some law schools in the common-law world have introduced into their curricula "general" subjects, such as "common law," in place of separate courses in contract and tort, or they require the student to write papers about issues that relate to several of the subjects studied.

No formal test is wholly satisfactory as a method of screening potential lawyers. The type used most widely, in which students write answers to questions in an examination hall, has been criticized for placing too much emphasis on memory. This criticism is met to some extent in many universities by allowing candidates to consult books and reference materials during the examination, thus bringing the test a little closer to what a lawyer will do when confronted with a real problem. Another objection is that testing creates a situation of stress, in which a candidate does not necessarily demonstrate how he has benefited from his legal education, and also one in which the skill demonstrated in the examination hall is not all the skill required of a lawyer. In particular, the examination does not test capacity for patient research or the capacity for oral argument, which requires theses and oral examinations. Examinations to be done outside orthodox examination halls have thus been proposed.

Some universities in the United States, Great Britain, and the Commonwealth countries require one or more long essays or a short thesis or research paper as part of the work for a first degree in law (as opposed to the more substantial dissertation, or thesis, for a postgraduate law degree). This is commonly written during the final year with no restriction on the resources employed. A thesis in the last year of study is required in some civil-law countries. Credit is also sometimes given for articles or notes published by students in law reviews. Such student publishing is more common in the United States than elsewhere, partly because most U.S. law schools have their own legal journals and partly because American law students are nearly always college graduates. Such student work also enhances prospects of employment, particularly if the student becomes an editor of the law review.

Oral examinations are the rule in some countries, such as Italy. In the United States oral examinations are rare. French universities typically use both written and oral examinations. Some British and overseas Commonwealth universities hold oral examinations to confirm or resolve doubtful results on written papers or as a prerequisite to the award of first class honours. In Italy, where a law student must present a thesis after passing his other examinations, the thesis must be orally defended before examiners. The German law student, after passing his written examination, has an oral one. In Japan, for professional qualification at the Legal Training and Research Institute (see below), there is an oral examination in each of the compulsory subjects after the written examination has been passed.

Qualifications for practice. *Common-law countries.* In England and Wales practicing lawyers must be either barristers (advocates and consultants) or solicitors (general legal advisers dealing with all kinds of legal business out of court and advocates in some of the lower courts). The former are organized in four Inns of Court (Lincoln's Inn, Inner Temple, Middle Temple, Gray's Inn) under the discipline of the Senate of the Inns of Court; the latter are under the jurisdiction of the Law Society. It is not necessary to hold a law degree or any university degree to qualify for the profession of law, but such a degree is usual. To become a barrister a candidate must pass a two-part examination in legal subjects, but university graduates may obtain partial or total exemption from the first part, depending on their degrees. A barrister's preparation also includes practical courses and a period of pupilage administered under the authority of the Senate of the Inns of Court. A barrister may not practice at all until he has undergone six months of pupilage in chambers and may not practice independently until he has been a pupil for a year. Pupilage causes some difficulty, partly because of the cost but mainly because of the increasing shortage of places in chambers. To qualify as a solicitor, the normal course is to serve as an articled clerk for two years and also pass law examinations in two parts. In Scotland and Ireland (both the republic of Ireland and Northern Ireland) there are similar requirements, though the arrangements differ in detail.

Preparation for a barrister

In the United States admission to the bar qualifies one for all types of legal work. The only formal requirements are the passing of state bar examinations after graduating from a law school; in a few states the law degree alone is sufficient.

In both England and the United States, as in many other common-law countries, becoming a judge or magistrate is a promotion (by appointment or election) from the ranks of the bar, and there is no special training for the exercise of judicial functions. But in some other common-law countries, especially in Africa and Asia, a newly qualified lawyer may enter the government legal service and find himself appointed in a short time to a junior magistracy. Even in these countries there is generally no special training for the job of adjudicating.

Civil-law countries. In continental European countries the qualifications to practice law typically depend on which of the various branches of the profession the university law graduate wishes to enter. Some countries place more emphasis on apprenticeship and others on examination. In France, for example, a legal practitioner may be an advocate, an *avoué,* a notary, or a judge. Each receives a different training, but all normally have gone through third- and fourth-year law degree courses. The advocate (roughly corresponding to the English barrister) must pass a bar examination and then serve as a probationary lawyer for three years, during which he takes further course work as well as acquiring practical experience. The *avoué* (something of a cross between a junior barrister and a senior solicitor) serves a period of articled clerkship and undergoes a professional examination by practicing lawyers. The notary (who does the noncontentious work performed in England by a solicitor) need not be a university graduate and can be a product of a professional school. His period of training lasts at least three years in a notary's office. He also takes a professional examination and if successful must wait for a vacancy since there is a limited number of notarial offices established by law.

In Germany the graduate in law who seeks a legal career must embark upon a period of practical training as a *Referendar.* This is a uniform program involving two years of practical work in the courts, in the office of a lawyer in private practice, in the office of a public prosecutor, in the civil service, and sometimes in the legal department of a commercial concern. Upon its completion, the graduate must pass a state examination (*Assessorexamen*).

A somewhat similar procedure is followed in Japan. Law graduates who seek careers as judges, public procurators, or lawyers in private practice must (with the exception of summary court magistrates and assistant procurators) pass the National Law Examination for entrance to the Legal Training and Research Institute. This is an organ of the Supreme Court. Like his German counterpart, the *Referendar* training to become an *Assessor,* the Japanese student at the institute is paid by the state. The period of training at the institute lasts two years. The bulk of the work consists of practical exercises and discussions, lectures on legal topics, and visits to institutions of concern to lawyers (such as prisons). The training is uniform, leads to a single examination, and qualifies the graduate for any branch of legal practice.

Judicial training

In some countries, such as France and Spain, there are special schools for training judges. In others, such as Germany and the Nordic countries, judicial training is acquired in the post-law-school practical internship period. In Germany, for example, a law graduate may be appointed to a lower court after completing the *Referendarzeit* and passing the Second State Examination. After serving a three-year probationary period, he becomes eligible for an appointment for life. In France the first step to becoming a judge is to pass an annual competitive examination for which students prepare by taking a special program in their last year of law studies. Successful candidates then must undergo 28 months of training consisting of a period of formal study at the National School of the Judiciary in Bordeaux, followed by a series of short practical internships in such settings as police departments, law offices, prisons, and the Ministry of Justice in Paris. This training culminates in a judicial apprenticeship, during which the future judge participates on a daily basis in all the activities of a variety of courts. Upon completion of their training period, the students are ranked on the basis of their grades and the evaluations of supervisors and are then assigned to their first positions in the judicial system. Since the administrative law courts in France are not part of the judiciary but rather of the administration, most judges for these courts are drawn not from the lawyers trained in the National School of the Judiciary but from the civil servants trained in the National School of Administration.

LEVELS OF STUDY

Law degrees are undergraduate degrees in most countries. The student embarks upon the study of law at a university at about the age of 18. In France the universities offer a course of two years in duration that may be taken by anyone who has completed his secondary education. High marks in this entitle the candidate to enroll for the *licence-en-droit,* which is given at the end of the third year of study. Successful completion of a fourth year leads to a *maîtrise-en-droit,* which for all practical purposes has become the basic French law degree.

In the United States, by contrast, most law schools require the entrant to be a university graduate. Consequently, the U.S. student of law is generally in his early 20s. Other countries where legal education is organized on a graduate level include India and Pakistan.

University law schools in many countries accept all candidates with a certain level of prelegal education. One drawback of this open admissions system is a substantial failure rate in examinations. In countries where candidates are screened before admission to law school there is less attrition. In England, for example, each university imposes a quota on entry to its law school and selects among candidates on the basis, usually, of academic performance. In the United States candidates are selected on the basis of academic performance and the results of a test designed to demonstrate aptitude for the study of law. In both the United States and England entry requirements vary according to the prestige of the law school.

Higher degrees in law

Most countries also provide for higher degrees in law. In common-law countries there is usually a series of steps, ascending through a degree of master of laws to a doctorate or senior doctorate. In civil-law countries it is normal to go straight from a first degree to a doctorate. Master's degrees are, as a rule, based on advanced examination after courses of instruction, though sometimes they are awarded for research or for a combination of examination and dissertation. Doctorates are awarded for theses expounding the results of original research and senior doctorates for published contributions to scholarship in the subject. In many countries there are also specialized postgraduate diplomas or certificates in particular subjects.

TRENDS IN LEGAL EDUCATION

Modern legal education is expanding both in quantity and scope, and formal university legal education has become dominant everywhere. The opportunities for university and professional education in law increased greatly after World War II. In England and Wales, for example, where before the war only the universities of Cambridge, Oxford, and London produced significant numbers of law graduates, there are now more than 50 academic law departments. A second wave of expansion took place, starting in the 1960s, when the numbers of law students, law instructors, and institutions teaching law grew dramatically almost everywhere. A particularly noteworthy development has been the increase in women law students, once a rarity but now constituting 30 to 40 percent of law students in most countries. The number of students enrolled in accredited law schools in the United States tripled between 1961 and 1980; thereafter demand for legal education began to level off and decline somewhat. A large increase in the teaching of law has occurred in Africa, where newly independent countries have established universities and professional schools concentrating on local laws and practice. Many governments have made provision, or greater provision, for the financial support of students, and legal education has been opened to a larger cross section of society in many places. The children of middle-class parents nevertheless continue to predominate in the field.

Since the late 1960s universities in several civil-law countries have departed from the rigidity of prescribed syllabi to allow a greater range of student choice in selecting subjects. In most countries, more attention is paid than formerly to foreign legal systems, transnational law, and to comparative law. In some countries nonlegal subjects have long been part of the syllabus; in others where law as a first-degree specialization has hitherto comprised only law studies, there has been a tendency to include nonlegal studies, joint courses in law and social sciences, or a more sociological approach to law.

Legal education has always had the problem of reconciling its aim of teaching law as one of the academic disciplines with its goal of preparing persons to become members of a profession. Most law schools are trying to find a middle path between being mere trade schools or citadels of pure theory. The criticism is often made that these efforts result in a type of education that is not practical enough to be really useful in resolving day-to-day legal problems but yet not as rigorously theoretical as a truly academic discipline ought to be. (L.A.S./M.A.Gl.)

BIBLIOGRAPHY

Legal profession: J.H. WIGMORE, *Panorama of the World's Legal Systems,* 3 vol. (1928, reissued in 1 vol., 1936), a comparative historical survey; DERK BODDE and CLARENCE MORRIS, *Law in Imperial China* (1967, reprinted 1973), a comprehensive account closely based on original sources; ROBERT J. BONNER, *Lawyers and Litigants in Ancient Athens: The Genesis of the Legal Profession* (1927, reprinted 1969); FRITZ SCHULZ, *History of Roman Legal Science* (1946, reprinted 1967), one of the few books on Roman law concentrating on the role of lawyers; HAROLD DEXTER HAZELTINE, "Roman and Canon Law in the Middle Ages," in *The Cambridge Medieval History,* vol. 5, ch. 21, pp. 697–764 (1926, reprinted 1968), a classic account; HERMAN J. COHEN, *History of the English Bar and Attornatus to 1450* (1929, reprinted 1967), a classic account of the history of the legal profession in the Middle Ages; MICHAEL BIRKS, *Gentlemen of the Law* (1960), a definitive work on the English solicitor; R.E. MEGARRY, *Lawyer and Litigant in England* (1962); and BRIAN ABEL-SMITH and ROBERT STEVENS, *Lawyers and the Courts: A Sociological Study of the English Legal System, 1750–1965* (1967), two opposing views of the legal profession in England; SIR FRED PHILLIPS, *The Evolving Legal Profession in the Commonwealth* (1978), a comparative survey of the legal profession, ethics, and education in countries with legal systems modeled on the English; DIETRICH RÜSCHEMEYER, *Lawyers and Their Society* (1973), a comparative study of the profession in West Germany and the United States; MARY ANN GLENDON, MICHAEL WALLACE GORDON, and CHRISTOPHER OSAKWE, *Comparative Legal Traditions* (1985), an introduction to the study of Romano-Germanic, English, and socialist law with material on the legal profession in those systems; C.J. DIAS *et al.* (eds.), *Lawyers in the Third World: Comparative and Developmental Perspectives* (1981); HIDEO TANAKA (ed.), *The Japanese Legal System* (1976, reprinted 1982).

Legal ethics: On the social nature and function of professional ethics, see ROBERT M. MACIVER, "The Social Significance of Professional Ethics," *Ann. Am. Acad. Polit. Soc. Sci.,* 297:118–124 (January 1955); and EMILE DURKHEIM, *Professional Ethics and Civic Morals,* trans. by CORNELIA BROOKFIELD (1957, reprinted 1983; orginally published in French, 1950). The leading texts on the ethics of the legal profession in the United States are HENRY S. DRINKER, *Legal Ethics* (1953, reprinted 1980); and L. RAY PATTERSON, *Legal Ethics: The Law of Professional Responsibility,* 2nd ed. (1984). The legal ethics of the Canadian legal profession are dealt with in MARK M. ORKIN, *Legal Ethics: A Study of Professional Conduct* (1957). For the English legal profession, see SIR THOMAS LUND, *Professional Ethics* (1970); and SIR WILLIAM BOULTON, *A Guide to Conduct and Etiquette at the Bar of England and Wales,* 6th ed. (1975). Information on legal ethics in other countries may be found in PIERRE G. LEPAULLE, "Law Practice in France," *Columbia Law Review,* 50:945–958 (1950); MAURO CAPPELLETTI, JOHN HENRY MERRYMAN, and JOSEPH M. PERILLO, *The Italian Legal System* (1967); TAKAAKI HATTORI, "The Legal Profession in Japan: Its Historical Development and Present State," in ARTHUR T. VON MEHREN (ed.), *Law in Japan* (1963); and K. OHIRA and G.N. STEVENS, "Admission to the Bar, Disbarment and Disqualification of Lawyers in Japan and the United States: A Comparative Study," *Washington Law Review,* 38:22–27 (1963). On the subject of contingent fees, see F.B. MACKINNON, *Contingent Fees for Legal Services: A Study of Professional Economics and Responsibilities* (1964). Specialization in the U.S. legal profession is discussed in BARLOW F. CHRISTENSEN, *Specialization* (1967). The same author discusses group practice in *Group Legal Services* (1967); his *Lawyers for People of Moderate Means: Some Problems of Availability of Legal Services* (1970), covers the subject of fees, specialization, group services, and advertising and solicitation.

Legal education: For England and Wales, much interesting information may be found in the GREAT BRITAIN. COMMITTEE ON LEGAL EDUCATION, *Report* (1971), which also contains a survey of legal education in Australia, Canada, France, Germany, India, Italy, The Netherlands, New Zealand, Nigeria, Scotland, South Africa, Sweden, and the United States. In Great Britain the journal *Legal Studies* (three times a year) publishes articles on legal education, chiefly but not exclusively relating to the United Kingdom; a similar periodical in the United States is the *Journal of Legal Education* (quarterly). ROBERT STEVENS, *Law School: Legal Education in America from the 1850s to the 1980s* (1983), is a comprehensive treatment of the subject. Legal education in parts of Asia is discussed in REGIONAL CONFERENCE ON LEGAL EDUCATION, 1962, SINGAPORE, *Report,* ed. by S.P. KHETARPAL (1964); HAKARU ABE, "Education of the Legal Profession in Japan," in ARTHUR T. VON MEHREN (ed.), *op. cit.;* and JAY MURPHY, *Legal Education in a Developing Nation: The Korea Experience* (1967). For Africa, see JOHN S. BAINBRIDGE, *The Study and Teaching of Law in Africa* (1972).

(G.S./M.E.P./L.A.S./M.A.Gl.)

Animal Learning

That animals can learn seems to go without saying. The cat that runs to its food dish when it hears the sound of the cupboard opening; the rat that solves a maze in the laboratory; the bird that acquires the song of its species—these and many other common examples demonstrate that animals can learn. Yet what is meant by saying that animals can learn? What, in other words, is learning? This question proves exceedingly difficult to answer, and, in fact, some theorists propose that no single, all-encompassing definition of learning is at all possible. Moreover, a moment's reflection yields the realization that there exist different kinds of learning. The learning of number concepts, for example, surely seems to be of a different nature than the learning of the association between the sound of a cupboard door and the receipt of food. To explore animal learning, then, this article first considers what learning is and is not and then examines in detail some of the specialized types of learning that occur in animals.

The article is divided into the following sections:

The general nature of learning

Many animals live out their lives following fixed and apparently unvarying routines. Among numerous species of solitary insects, for example, the life cycle consists of the following unvarying events: the females lay their eggs on a particular plant or captured prey; the newly hatched larvae immediately start eating and then follow a standard sequence of developmental stages; the adults recognize appropriate mates by a set of fixed signs, perform a fixed sequence of mating responses, provision their eggs with suitable nourishment, and finally die before the next generation hatches. The same unchanging sequence is repeated generation after generation. And it is, of course, eminently successful. The same set of responses is invariably elicited by the same set of stimuli, because those responses were, and continue to be, adaptive. Where circumstances do not change, there is little need for an animal's behaviour to change. Even many aspects of the behaviour of mammals show a similar fixity. A dog withdraws its foot if it is pricked and a young child his hand if burned; both people and rabbits blink whenever an object is moved rapidly toward their eyes; the feeding behaviour of young infants of virtually all mammalian species consists of sucking elicited by contact with the lips.

Whenever the same response is always appropriate in a particular circumstance, there is little reason why an animal should need to learn what to do in that circumstance. But the world is not always so stable a place. The food supply that was plentiful yesterday may be exhausted today, and the foraging animal that always returns to the same spot will starve to death. Moreover, a particular food supply may be temporarily depleted but will be replenished if left long enough; the successful forager needs to remember where the supply was and when it was last visited, so as to time a return to advantage. In other words, circumstances may change, and the same response is not always appropriate to the same stimuli. Knowing what behaviour is appropriate may depend, therefore, on keeping track of past events.

POSSIBLE EXPLANATIONS OF BEHAVIORAL CHANGES

If an animal's behaviour toward a particular stimulus changes, one must look for an explanation of that change. One possible explanation is that the change is due to learning, but there are numerous other possibilities. If a definition of learning is to be provided, that definition must specify when to attribute the change to learning, and when to other causes.

At least two other major causes of behavioral change have been widely recognized. The first of these is motivation. A laboratory rat may pick up, chew, and swallow a pellet of food at one moment; half an hour later, after having eaten 20 grams of food, the rat will simply ignore any further pellets offered. Similarly, a male rat may mount and copulate with a receptive female introduced into his cage, but he will not repeat this pattern of behaviour endlessly even if offered the opportunity to do so. Some male territorial birds, such as chaffinches, will feed amicably beside other males at certain times of day or certain seasons of the year, but at other times they will launch an attack on any intruding male. In all these cases, it is more reasonable to attribute the change in behaviour not to anything the animal has learned but rather to a change in the creature's motivational state.

Changes due to motivation

It should not be thought, however, that just because all of these examples can be attributed to a single item—*i.e.*, motivation—that their detailed explanation will always be the same. The analysis of motivation is itself a large field of study, and it has proved to be more profitable to concentrate on the specific explanation of individual cases of changes in behaviour rather than to search for broad explanatory principles that end up being nearly vacuous. Nonetheless, it does seem possible to draw a contrast between motivational explanations for such changes and those that appeal to learning.

A second broad class of changes in behaviour can be attributed to maturation. We are inclined to ascribe the unfolding pattern of behaviour that emerges over the first few weeks of life to this ill-defined process. Newborn rat pups, for example, are relatively helpless; their eyes do not open for about two weeks, and their main sources of sensory input are probably touch and smell. As their sensory apparatus matures, the pups exhibit changed behavioral responses. The other obvious instance of a maturational change in behaviour is that which comes with sexual maturity: sexually mature adults of most species behave toward one another in ways quite different from those of

Changes due to maturation

younger members of the species. It is not only courtship and mating behaviour that change with sexual maturity; for instance, male puppies urinate in the same way as females, by squatting, and it is the onset of sexual maturity that produces the adult pattern of cocking a hind leg.

The concept of maturation is probably no better defined than that of motivation, and it is equally important to stress that it must cover a number of different processes. And, as with motivation, it is more profitable to analyze each case in detail, in order to uncover the precise mechanisms involved, than it is simply to label a change as an example of maturation. Indeed, at the level of physiological process, it seems probable that both motivational and maturational changes are often due to alterations in the hormonal state of the animal, and the distinction between the two is largely one between the unidirectional nature of the change in the case of maturation contrasted with the cyclical change common to short-term motivational states.

But how are these changes discerned from those that might be ascribed to learning? In many cases, of course, the answer is because a precise causal explanation has been provided: a great deal is known, at a physiological level, about the changes in brain and body associated with the motivational states of hunger and thirst. Even without any such detailed knowledge of the underlying mechanisms, it is possible to insist that certain changes in behaviour be attributed to motivation rather than to learning if the opportunity to learn anything relevant was lacking and the opportunity for a motivational change was present. If, for example, an animal that has been deprived of food for a long time behaves in one way toward food-related stimuli, but some hours later, after having been given ample opportunity to eat, it behaves differently toward those stimuli, the obvious interpretation is a motivational one. This interpretation would be strengthened if the animal had not come into contact with these stimuli during the intervening period and had been given, as far as one could judge, no other opportunity to learn anything about them. Learning, in other words, depends on certain kinds of opportunity, and a definition of learning may well turn out to be no more than a specification of the particular set of opportunities and experiences that produce it.

CIRCUMSTANCES THAT PRODUCE LEARNING

A particular change in behaviour is attributed to learning, then, because it is possible to specify the set of circumstances that produced it. What are those circumstances? It is common to claim that learning depends on practice. (An older generation of experimental psychologists would have claimed that it depended on "reinforced" practice.) This definition can be misleading, however, if it causes one to attribute to learning all behavioral changes that follow what appears to be practice. In other words, it is not enough to show that an animal appeared to engage in practice and its behaviour subsequently changed. A temporal correlation of this sort does not establish a causal connection. Young birds, for example, are unable to fly, and their first attempts at flight are clumsy and ill-coordinated. Casual observation suggests that young birds improve with practice, gradually perfecting the set of skills they display as adults, but experimental analysis suggests that this practice may be unnecessary. Young birds have been brought up under restricted conditions that completely prevented their flying. When released at the age at which normally reared birds fly proficiently, the experimental subjects flew—without practice—as successfully as those that had spent their time in trial flight. The development of the skill appears to depend more on the maturation of strength and agility than on specific practice.

The role of practice

The notion that learning depends on practice also seems unduly restrictive and is, perhaps, an unnecessary legacy of an earlier version of behaviourism. It is not obvious that an animal should actually have to engage in a particular form of behaviour in order that this pattern of behaviour should be affected by learning. In many cases, indeed, no such practice seems necessary. The young of many songbirds must, it is quite clear, learn their species-typical song. There are several aspects to this learning process, one of which may indeed involve practicing the song at the beginning of the young bird's second season. But another critical aspect is simply exposure to the adult song at some point during the autumn of the young bird's first year, at a time when the young bird does not practice singing at all. Deprived of such experience, chaffinches and song sparrows produce an extremely impoverished version of the adult song; some finches may develop a song more characteristic of another species if that is what they heard during this period of their life. There are numerous other examples where learning appears to depend more on the opportunity to observe than on the opportunity for practice.

Broad versus narrow definitions of learning

This suggests that the definition of learning will have to refer to changes in behaviour that are attributable to particular kinds of experience. The danger now is that, as with motivation and maturation, the definition of learning will be so broad and vague as to be useless. As in those cases, it may be more profitable to concentrate on more detailed analysis of particular instances of learning. Such analysis has, for example, led to widespread agreement on the definition of classical conditioning, a particular type of learning whose study was pioneered by the Russian physiologist Ivan Petrovich Pavlov. In a typical experiment on classical conditioning, an experimenter might arrange a correlation between the ringing of a bell and the delivery of food to an animal. The animal predictably learns to direct food-related activity toward the sound of the bell. Analyses of such experiments have led to the definition of classical conditioning as a type of learning that occurs when there is a correlation between two stimuli and the animal's behaviour toward one of these stimuli changes in a predictable manner determined by the nature of the other. This definition, which will be expanded later in this article, is useful because it specifies both the circumstances responsible for learning (a temporal correlation between two stimuli) and the general way in which experience of those circumstances changes behaviour (the animal starts directing toward one stimulus responses that are related to those normally directed toward the other). Experimental psychologists and ethologists, however, have devised a tremendous range of procedures for studying learning in animals. The range and variety are such that it may be well-nigh impossible to formulate a meaningful definition of the circumstances that produce learning, for the definition either will be so restrictive that it clearly applies to only a fraction of the cases that should be regarded as instances of learning, or it will be so broad that it says nothing.

Rather than pursue any further the attempt to find an all-embracing, single definition of learning, it seems more useful to provide narrower definitions for particular cases, along the lines suggested above for classical conditioning. One consequence of this approach is that it may encourage the belief that learning consists of a large number of distinct processes that have nothing in common with one another. It is, of course, an open question as to whether this is true: it is certainly possible that, just as with the concept of motivation, the layman's concept of learning encompasses a large number of different cases whose underlying mechanisms are quite distinct. It is important not to prejudge this issue. Insistence on a single, global definition may well tend toward just such prejudgment by encouraging the belief that learning is a single, common process. To start by drawing some distinctions between types of learning does not rule out the possibility of seeing whether the various cases studied do have anything in common.

The interplay of theory and definition

In the final analysis, as is true of all scientific definitions, the definition of learning is a matter of theory. It has been said that a good scientific definition is the end product of good theory and experiment, not the starting point. Thus, there is a single process of learning if it turns out to be possible to devise a single theory that adequately accounts for the variety of cases in which learning is assumed to occur. Superficial appearances may be deceptive: just because the circumstances that produce learning in two cases, along with the consequences of that learning, appear quite different, it does not follow that the processes underlying learning are different. For instance, the phenomenon of filial imprinting, first seriously analyzed by the Austrian ethologist Konrad Lorenz, appears to be a highly specialized form of learning in which a newborn animal (*e.g.*, a chick, duckling, or gosling) rapidly learns to follow the first salient, moving object it sees. Normally this object will be the mother, but Lorenz discovered that the range of potential imprinting objects is large, extending from Lorenz himself to a bright red ball. There is no question but that some process of learning occurs here, and Lorenz assumed it to be highly specialized. Yet one theory seeks to explain imprinting in terms of simple classical conditioning. Whether or not the account of imprinting provided by this theory is correct, the point is made that how learning is defined and whether it is defined as a single, monolithic process or as many specialized processes are, in the end, questions of theory.

Types of learning

SIMPLE NONASSOCIATIVE LEARNING

When experimental psychologists speak of nonassociative learning, they are referring to those instances in which an animal's behaviour toward a stimulus changes in the absence of any apparent associated stimulus or event (such as a reward or punishment). Studies have identified two major forms of simple nonassociative learning, which are to some extent mirror images of one another: habituation and sensitization.

Habituation. A classic example of habituation is the following observation on the snail *Helix albolabris.* If the snail is moving along a wooden surface, it will immediately withdraw into its shell if the experimenter taps on the surface. It emerges after a pause, only to withdraw again if the tap is repeated. But continued repetition of the same tapping at regular intervals elicits a briefer and more perfunctory withdrawal response. Eventually, the stimulus, which initially elicited a clear-cut, immediate response, has no detectable effect on the snail's behaviour. Habituation has occurred.

Habituation can be defined in behavioral terms as a decline in responding to a repeatedly presented stimulus. As such, it is a very widespread phenomenon, one that can be observed in animals ranging from single-celled protozoans to humans. Most animals behave differently to novel and familiar stimuli: the former sometimes elicit startle responses, sometimes investigatory or exploratory responses; the latter often apparently are ignored. The suggestion that habituation is a simple form of learning, however, implies that it can be distinguished from some even simpler potential causes of this sort of change in behaviour. One reason why an animal might stop responding to a stimulus is that it no longer detects the stimulus; *i.e.*, some form of sensory adaptation might have occurred. Another potential cause is fatigue: perhaps some temporary refractory state is produced by repeated elicitation of the same response, making it impossible to perform that response again. Whether or not one would want to call either of these processes a form of learning is doubtful. But both behavioral and physiological evidence establishes that habituation cannot be explained in these terms.

Evidence that habituation is a form of learning

The critical behavioral evidence is that habituation can be disrupted by almost any change in the experimental conditions. If repeated presentation of one stimulus leads to habituation of a response, the same response can still be elicited by a different stimulus. Even if the experimenter presents a novel stimulus that does not itself elicit the response in question, its presentation may restore the response on the next trial in which the originally habituated stimulus is presented. This latter observation, usually referred to as an instance of dishabituation, seems to rule out any simple sensory adaptation; both observations rule out simple effector fatigue.

Neurophysiological analysis of habituation in various mollusks—for example, in the sea snail *Aplysia*—has confirmed that habituation need not depend on changes in the activity of sensory or motor neurons. In the case of *Aplysia*, researchers have studied the gill withdrawal reflex, a response that rapidly habituates to repeated stimulation of the snail's siphon or mantle shelf. But habituation still occurs even if it is elicited by direct, electrical stimulation of the motor nerve, bypassing the sensory receptors completely; and recording from the sensory nerve during normal habituation reveals no decline in its level of activity. These observations eliminate sensory adaptation as a possible cause of the animal's having ceased to respond to the stimulus. Effector fatigue can be ruled out by showing that direct stimulation of the motor neurons controlling the withdrawal response can still elicit a perfectly normal reaction even after the response has completely habituated. Research shows that habituation in *Aplysia* depends on changes in the activity of more central neurons. Repeated tactile stimulation of the siphon, leading to habituation of the withdrawal response, causes changes in the activity of the motor neurons innervating the response. Specifically, these motor neurons show a decline in excitatory postsynaptic potential, which is the electrical change that enables the nerve impulse to cross the gap (synaptic cleft) that separates one neuron in the pathway from the next. The decline in excitatory postsynaptic potential short-circuits the response. Moreover, the presentation of a novel stimulus, sufficient to dishabituate the behavioral response, restores the postsynaptic potential.

Habituation occurs even in animals without a central nervous system—probably in single-celled protozoans; certainly in animals such as the coelenterate *Hydra,* which have a diffuse nerve net and do not appear to be capable of associative learning. Among mammals, habituation of certain reflex responses can be observed even in "spinal" subjects, that is, those whose spinal cord has been severed from the brain. There can be little doubt, then, that habituation is not only widespread, but that it also can be a relatively simple phenomenon. There is, however, no guarantee that it is the same phenomenon wherever it appears. The waning response to a repeatedly presented stimulus admits of a number of different explanations. In principle, as we have already seen, it might be due to sensory adaptation, effector fatigue, or a more central neural change. These distinctions make rather little sense in the case of a single-celled animal. And one should not necessarily expect the habituation observed in a spinal mammal to involve precisely the same mechanisms as those responsible for comparable behavioral effects in an intact animal. Some psychologists have proposed theories of habituation that appeal to processes of classical conditioning. Such a theory is not likely to apply to the habituation observed in an animal that shows no capacity for classical conditioning.

Habituation is usually, as here, classified as an instance of simple, nonassociative learning. It is supposedly nonassociative because all that happens in the course of habituation is that a stimulus is repeatedly presented and the animal's behaviour changes; there is, on the face of it, no other event with which the stimulus can be associated. Habituation must therefore, it appears, be understood by reference to some change in the pathway between stimulus and response, and the work with *Aplysia* and other mollusks shows how this analysis may proceed at the physiological level. But if habituation is not always the same phenomenon, it is possible that different processes may underlie the habituation of the startle response to a loud noise in an intact mammal. And despite appearances to the contrary, those processes may involve some associative learning. One suggestion is that novel stimuli elicit a biphasic response: an initial increase in startle responses, which include components of emotion or anxiety, followed by a rebound in the opposite direction. Habituation occurs when the latter, rebound response becomes conditioned to the stimulus, occurring sooner and sooner with each repetition of the stimulus and thereby damping down and eventually canceling out the initial reaction. An alternative possibility is that long-term habituation depends on associating the repeatedly presented stimulus with the context in which it occurs, a suggestion that would explain why presentation of the stimulus in a different context sometimes leads to dishabituation.

Possible associative bases for habituation

The generality of habituation implies that this behavioral phenomenon has considerable adaptive significance; if true, it would be quite reasonable to expect that a number of different mechanisms might have evolved to produce the behavioral result. The adaptive value of habituation is not difficult to see. A novel stimulus may signify danger, and an animal should react to this stimulus either by withdrawing or at least by orienting toward it to see what will happen next. But if the same stimulus occurs again with no further consequence, it is probably safe: regular repetition of the same stimulus implies that it is part of the background, such as the waving of a branch in the wind or the shadow caused by a piece of seaweed floating with the waves. If the stimulus is not dangerous, time should not be wasted on it. Withdrawal, especially in the case of a snail into its shell, is a time-consuming effort, incompatible with such vital activities as searching for food. If it is important, therefore, for animals to be wary of novel stimuli, it is equally important that they should discriminate the novel and potentially dangerous from the familiar and probably safe.

Adaptive value of habituation

Sensitization. The effect of habituation is to eliminate unnecessary responses, but the main function of learning has usually been thought to be the production of new responses. Traditional psychological theories of learning have assumed that the learning of new patterns of behaviour comes about through the association of a new response with a particular stimulus. Consequently, psychologists usually have either ignored the possibility that nonassociative processes might be sufficient to increase the probability of a new response or regarded it as a nuisance that interferes with the measurement of associative changes. They have rarely treated it as a subject worthy of study in its own right.

This is unfortunate, for the nonassociative phenomenon of sensitization is probably fairly widespread, and it provides a simple means of acquiring adaptive behaviour. Sensitization is said to occur when the repeated presentation of a particular significant stimulus (such as food or electric shock) lowers the threshold for the elicitation of appropriate behaviour to the point where a second stimulus, not normally capable of calling forth that behaviour, now does so. A typical example is provided by the behaviour of the marine worm *Nereis.* If the worm is kept in a small tube and fed at regular intervals, it becomes progressively more likely to respond to any novel stimulus, such as a change in illumination, by exploratory, food-seeking movements toward the open end of the tube. If, on the other hand, the worm receives mild electric shocks at regular intervals, it becomes progressively more likely to respond to a novel stimulus by withdrawal.

Relationship of sensitization to habituation

The first point to note about sensitization is its relationship to habituation. Habituation refers to a decline in the probability of responding to a repeatedly presented stimulus. Sensitization, by contrast, refers to an increase in the probability that behaviour appropriate to a repeatedly presented stimulus will occur, even in response to another stimulus. Although these two outcomes cannot be observed simultaneously, it is quite possible that the same operation—repeated presentation of a stimulus—can simultaneously engage two different processes, one causing a decline in the probability of responding to that stimulus, the other causing an increase. Experimental analysis suggests that both processes are real and may be engaged in the same experiment, so that the observed change in behaviour actually results from a mixture of the two. Typically, the process of habituation wins out, and what is observed is an overall decline in responding. But a common finding in habituation experiments is that responding initially increases before declining; the implication is that the initial presentations of a stimulus result in more sensitization than habituation, while further presentations produce more habituation than sensitization. A second factor influencing the relative importance of the two processes is the intensity or significance of the stimulus. A weak stimulus, or one with little intrinsic biological significance, will show relatively rapid habituation and little or no initial sensitization. A stronger stimulus, especially one, such as food or shock, that has substantial significance to the animal, may show marked sensitization and relatively little habituation.

The second point about sensitization is that it may mimic the effect of associative learning or conditioning. As has been mentioned, in a classical conditioning experiment a neutral stimulus, such as a change in illumination, is paired with the delivery of a significant stimulus, such as food or shock. Repeated pairing causes the neutral stimulus to elicit responses initially called forth by the significant stimulus; for example, a change in illumination that has been associated with an electric shock would come to elicit retreat or withdrawal. But in the case of the worm *Nereis,* experiments demonstrate that the light would come to elicit this change in behaviour whether or not it had been paired with shock: all that is needed is sufficient exposure to the shock. To attribute the change in behaviour toward the light to its association with food or shock, one must show that this change is greater than that which would have resulted from sensitization alone.

Relationship of sensitization to associative learning

The physiological processes underlying sensitization, like those underlying habituation, have been analyzed in experiments on such invertebrate species as *Aplysia.* Not surprisingly, the mechanisms involved appear to mirror one another. Whereas habituation is correlated with a decline in postsynaptic potentials, sensitization is correlated with an increase in the magnitude of postsynaptic potentials at the same locus.

Although sensitization has often been treated as a nuisance whose effects must be controlled in studies of habituation or associative learning, it remains a process worthy of study in its own right, for the behavioral changes it produces can have significant adaptive value. Without requiring the presumably more complex neural machinery necessary to subserve associative learning, sensitization enables animals to respond to local variations in the occurrence of significant events. If an animal's sources of food tend to occur together (that is, they are not distributed randomly in time or space), then it pays that animal, having once found food, to continue to behave in a food-gathering manner. Conversely, the animal that is increasingly wary after exposure to danger will have a better chance of evading a lurking predator. Sensitization thus enables an animal to take advantage of statistical regularities in the occurrence of significant events, without requiring it to detect other events that predict the significant ones. No doubt, further advantage accrues to the animal that can perform such calculations, for associative learning provides a powerful means of predicting the future. But there can be equally little doubt that such a process requires a more elaborate nervous system.

Adaptive value of sensitization

ASSOCIATIVE LEARNING: CONDITIONING

The study of animal learning in the laboratory has long been dominated by experiments on conditioning. This domination has been resisted by critics, who complain that conditioning experiments are narrow, artificial, and trivial, and, as such, miss the point of what animals are adapted to learn. From the critics' point of view, one unfortunate effect of their attacks has been the progressive refinement and elaboration of the theory of conditioning to the point where it can often explain the exceptions to which they drew attention. This is not to insist that associative learning is the sole, or even the most important, form of learning in vertebrates, but rather to introduce the idea that the processes underlying conditioning may be more interesting than older theories and an earlier generation of textbooks suggested.

Classical and instrumental conditioning. Pavlov was not the first scientist to study learning in animals, but he was the first to do so in an orderly and systematic way, using a standard series of techniques and a standard terminology to describe his experiments and their results. In the course of his work on the digestive system of the dog, Pavlov had found that salivary secretion was elicited not only by placing food in the dog's mouth but also by the sight and smell of food and even by the sight and sound of the technician who usually provided the food. Anyone who has prepared food for his pet dog will not be surprised by Pavlov's discovery: in a dozen different ways, including excited panting and jumping, as well as profuse salivation, the dog shows that it recognizes the familiar precursors of the daily meal. For Pavlov, at first, these "psychic secretions" merely interfered with the planned study of the digestive system. But he then saw that he had a tool for the objective study of something even more interesting: how animals learn. From about 1898 until 1930, Pavlov occupied himself with the study of this subject.

Pavlov's experiments

Pavlov's experiments on conditioning employed a standard, simple procedure. A hungry dog was restrained on a stand and every few minutes was given some dry meat powder, an event signaled by an arbitrary stimulus, such as the ticking of a metronome. The food itself elicited copious salivation, but, after a few trials, the ticking of the metronome, which regularly preceded the delivery of food, also elicited salivation. In Pavlov's terminology, the food is an unconditional stimulus, because it invariably (unconditionally) elicits salivation, which is termed an unconditional response. The ticking of the metronome is a conditional stimulus, because its ability to elicit salivation (now a conditional response when it occurs in reaction to the conditional stimulus alone) is conditional on a particular set of experiences. The elicitation of the conditional response by the conditional stimulus is termed a conditional reflex, the occurrence of which is reinforced by the presentation of the unconditional stimulus (food). In the absence of food, repeated presentation of the conditional stimulus alone will result in the gradual disappearance, or extinction, of its conditional response. In translation from the Russian, the terms "conditional" and "unconditional" became "conditioned" and "unconditioned," and the verb "to condition" was soon introduced to describe the experimental activity.

Thorndike's experiments

To the American psychologist Edward L. Thorndike must go the credit for initiating the study of instrumental conditioning. Thorndike began his studies as a young research student, at about the time that Pavlov—already 50 years old and with an eminent body of research behind him—was starting his work on classical conditioning. Thorndike's typical experiment involved placing a cat inside a "puzzle box," an apparatus from which the animal could escape and obtain food only by pressing a panel, opening a catch, or pulling on a loop of string. Thorndike measured the speed with which the cat gained its release from the box on successive trials. He observed that on early trials the animal would behave aimlessly or even frantically, stumbling on the correct response purely by chance; with repeated trials, however, the cat eventually would execute this response efficiently within a few seconds of being placed in the box.

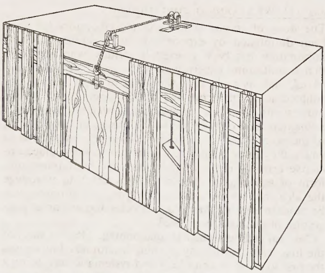


Figure 1: A Thorndike puzzle box. A cat placed within this apparatus could obtain its release by pressing a pedal inside the box.

Skinner's experiments

Thorndike's procedures were greatly refined by another U.S. psychologist, B.F. Skinner. Skinner delivered food to the animal inside the box via some automatic delivery device and could thus record the probability or rate at which the animal performed the designated response over long periods of time without having to handle the animal. He also adopted some of Pavlov's terminology, referring to his procedure as instrumental, or operant, conditioning; to the food reward as a reinforcer of conditioning; and to the decline in responding when the reward was no longer available as extinction. In Skinner's original experiments, a laboratory rat had to press a small lever protruding from one wall of the box in order to obtain a pellet of food. Subsequently, the "Skinner box" was adapted for use with pigeons, who were required to peck at a small, illuminated disk on one wall of the box in order to obtain some grain.

In experiments on both classical conditioning and instrumental conditioning, the experimenter arranges a temporal relation between two events. In Pavlov's experiment the food was always preceded by the conditional stimulus; in Skinner's original experiment the delivery of food was always preceded by the rat's pressing the lever. Conditioning, or associative learning, is inferred if the animal's behaviour changes in certain ways and if that change can be attributed to the temporal relationship between these events. If the dog started salivating to the ticking of a metronome just because it had recently received food, rather than because the delivery of food had been signaled by the metronome, this should be regarded as an instance of sensitization rather than associative learning. One of the simplest ways of establishing that the change in behaviour results from the temporal relationship between the conditional stimulus and the unconditional stimulus in a classical experiment, or between the response and the reinforcer in an instrumental case, is to impose a delay between the two. A gap of even a few seconds between the rat's pressing the lever and the delivery of food will seriously interfere with the animal's ability to learn the connection. And although in some classical experiments evidence of conditioning can be found in spite of relatively long gaps between the conditional stimulus and the unconditional stimulus, increasing this interval beyond a certain point invariably causes a decline in conditioning.

Laws of associative learning. The temporal relation between the conditional stimulus and the unconditional stimulus, or between the response and the reinforcer, was for a long time regarded as the primary determinant of conditioning. Conditioning is certainly a matter of associating temporally related events, but temporal contiguity is only one of several factors—and probably not the most important—that influences conditioning. A variety of experiments have shown that classical conditioning will occur only if the conditioned stimulus is the best predictor of the occurrence of the unconditional stimulus. In other words, it is the correlation between two events, just as much as their temporal contiguity, that establishes an association between them. A pigeon, for example, will learn by classical conditioning to peck an illuminated disk in a Skinner box if, whenever the disk is illuminated, food is delivered. This temporal relationship between the light and food can be preserved intact, but if the experimenter now arranges that food is equally available at other times (when the light is not on), the pigeon will not peck at the illuminated disk. Delivering food at other times destroys the correlation between light and food (although leaving the temporal relationship untouched) and abolishes conditioning.

Conditioning occurs to the best predictor

Although some conditioning will occur when the conditional stimulus is not perfectly correlated with the delivery of food (perhaps because on a proportion of trials the conditional stimulus is presented alone without food) or when the temporal relationship is less than perfect (there is a gap between the conditional stimulus and the delivery of food), this conditioning is abolished if the experimenter ensures that there is some better predictor always available. If a dog is conditioned to the ticking of a metronome paired with the delivery of food, the animal will salivate in response to the metronome even if the food is presented in no more than 50 percent of the trials. If, however, a light is illuminated on those trials when the metronome is accompanied by food, and not on the remaining 50 percent of the trials, the dog will become conditioned to the light and not to the metronome. Similarly, a pigeon will learn to peck at a disk illuminated with red light even if a gap of several seconds separates this response from the delivery of food. But if, during this interval, after the red light has been turned off and before food is delivered,

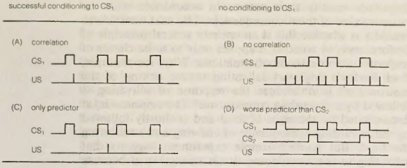


Figure 2: Effects on conditioning of various relationships between a conditioned stimulus (CS_1) and an unconditioned stimulus (US). (A) and (B) show that temporal contiguity of CS_1 and the US is not by itself sufficient to ensure conditioning; (C) and (D) demonstrate that conditioning occurs to the best predictor of the US (see text).

a green light is turned on, the pigeon will never learn to peck at the red light. It is as though the pigeon attributes the occurrence of food to the most recent potential cause (now the green light rather than the red), and the dog attributes food to the stimulus best correlated with its delivery (the light rather than the metronome). Conditioning, in other words, occurs selectively to better predictors of reinforcement at the expense of worse predictors. This same principle explains the earlier observation of the role of correlation in general. The pigeon will not associate the illumination of the disk with food if food is equally probable both when the light is on and when it is switched off; from the pigeon's point of view, food occurs whenever the animal is placed in the Skinner box. The illumination of the light signals no increase in the probability of food, and the best predictor of food is the mere fact of being in the Skinner box.

Temporal contiguity, therefore, is not necessarily the most important factor in successful conditioning. Moreover, there is yet another factor that should be stressed. It will hardly have escaped the reader's attention that there is an astonishing artificiality to the typical conditioning experiment conducted by Pavlov or Skinner. An animal is placed in a bare, confined space; lights are flashed on and off; the animal is permitted to operate some mechanical contrivance; some meat powder or a pellet of food is delivered. How could one possibly suppose that the ways in which animals learn anything of importance in the real world will be illuminated by this contrived and restrictive kind of experiment? This question raises large issues, some of which will recur at later points in this article. But one point should be acknowledged right away: the more restricted the range of experimental manipulations employed, the greater the chance that the investigator will completely miss important principles. Experiments with lights and metronomes failed to reveal the following important principle of conditioning: animals appear to have built-in biases toward associating some classes of stimuli with certain classes of consequences. The most dramatic instance of this principle is provided by conditioned food aversions. If rats eat some novel-flavoured substance and shortly thereafter are made mildly ill (for example, by an injection of a drug such as apomorphine or lithium chloride), they afterward will show a marked aversion to the novel food. Because they will show an aversion even though an interval of several minutes, or sometimes even hours, intervenes between eating the food and the onset of the illness, there has been some question as to whether this should be regarded as an instance of conditioning at all. But the parallels between food aversions and other forms of conditioning are so extensive that it is hard to believe that some common processes are not involved. And there is no question but that the length of the interval is important; other things being equal, rats will form a stronger aversion to a food they have eaten recently than to one they have eaten several hours earlier.

Built-in biases toward certain associations

The most interesting feature of such aversions is that they are, by and large, confined to foods. If rats suffer the unpleasant experience of being made ill, they are not likely to show an aversion to anything other than a novel-tasting food or drink they have recently ingested. As in other forms of conditioning, the novelty of the potential conditional stimulus is important. Rats will not show any marked aversion to a thoroughly familiar diet unless the experience of illness is repeatedly induced shortly after eating the daily ration, just as, in Pavlov's experiments, conditioning will proceed only slowly to the ticking of a metronome if the dog has heard this sound repeatedly before. The more striking restriction, however, is that it is the taste of the food or drink that is associated with illness. If rats drink plain tap water before being made ill, they will show little aversion to tap water (since there is no novelty here). But even if a novel buzzer is sounded while they are drinking and they are then made ill, they will not associate the buzzer with the illness. This is certainly not because rats are unable to associate the buzzer with an aversive consequence. If drinking water while the buzzer is sounded produces a mild electric shock, rats will rapidly learn to stop drinking whenever they hear the buzzer. In this case it is the flavour of the water that rats find difficult to associate with the shock; punishing rats with a mild shock whenever they drink sugar-flavoured water has little effect on their tendency to drink sugar-flavoured water. The flavour of food or drink is readily associated with subsequent illness, but only poorly associated with other painful consequences. Conversely, an external stimulus such as a buzzer or flashing light, which is readily established as a signal for shock, is only with great difficulty associated with illness. These relationships are summarized in the Table.

Occurrence of Conditioning to Certain Combinations of Conditional Stimulus and Unconditional Stimulus

conditional stimulus	unconditional stimulus	outcome
Water with novel flavour	Lithium-induced illness	Conditioned aversion to flavour
Water with novel flavour	Mild shock	No aversion to flavour
Water plus buzzer	Lithium-induced illness	No aversion to buzzer
Water plus buzzer	Mild shock	Conditioned aversion to buzzer

The full explanation of this finding remains uncertain. It is known that even very young rats show such selectivity, so it cannot depend solely on any prior experience. What is easy to see is that this behaviour makes biological sense. Internal malaise, such as that caused in the psychologist's experiment by an injection of lithium, will in the real world usually be a consequence of eating spoiled or poisonous food or of drinking tainted water. The most reliable sign of such food or drink will be its taste, and animals predisposed to associate the taste of what they have ingested with subsequent illness are likely to be better equipped to avoid potentially harmful food in the future. On the other hand, painful injury, mimicked in the laboratory by a brief electric shock, is hardly likely to be a consequence of eating food of a particular flavour; it will usually be caused by external circumstances, such as contact with a sharp or very hot object or a narrow escape from a predator. The natural suggestion is that the function of conditioning is to enable animals to find out what causes certain events of biological significance. If this is so, a built-in bias toward associating certain classes of events together makes adaptive sense. Conditioning is not just a matter of associating two events because one happens to follow the other; it is more profitably seen as the process whereby animals discover the most probable causes of events of consequence to themselves.

Adaptive value of biases

Laws of performance. Conditioning could have no function at all, however, if it did not involve changes in an animal's behaviour. Nor could scientists infer that conditioning has occurred unless they could observe, at some point, a change in an animal's behaviour attributable to certain conjunctions of events. So, although conditioning may involve the formation of associations between events or the attribution of particular events to their most probable antecedent causes, it must also include some mechanisms for translating these associations into changes in behaviour.

For an earlier generation of behaviourists, the fundamental fact about conditioning was precisely that it changed behaviour, and the theories they advanced were determined by this fact. The description of conditioning as the establishment of a new response to a stimulus that had not previously elicited that response naturally suggested that conditioning was a matter of forming new stimulus–response connections. This conceptualization led to the development of the stimulus–response theory, variations of which long provided the dominant account of conditioning. One version of the stimulus–response theory suggested that the mere occurrence of a new response to a given stimulus, as when Pavlov's dog started salivating shortly after the metronome had started ticking, is in itself sufficient to strengthen the connection between the two. Thorndike, however, argued that the probability that a particular stimulus will repeatedly elicit a particular response depends on the perceived consequences of this response. According to this view, new stimulus–response connections are strengthened only if the response is followed by certain kinds of consequences.

Stimulus–response theory

There are several questions raised here, and it is important to keep them distinct. One is whether responses are sometimes (or even always) modified by their consequences. Although denied by some theorists, their denial seems distinctly paradoxical. A rat whose presses on a lever are followed by the delivery of a food pellet will press the lever again; if the only consequence of pressing the lever is the delivery of a painful shock, the rat will desist from this action. Thorndike's law of effect—which stated that a behaviour followed by a satisfactory result was most likely to become an established response to a particular stimulus—was intended to summarize these observations, and it is surely an inescapable feature of understanding how and why humans and other animals behave. In keeping with this understanding, parents reward children for good behaviour and punish them for bad. When this fails to produce the desired behaviour, we are inclined to argue that the child is finding other sources of reward or does not find the intended punishment particularly unpleasant, or that the parents' behaviour is hopelessly inconsistent. We are far less likely to question the assumption that, other things being equal, people (and other animals) repeat actions that have desirable consequences and avoid repeating those that have undesirable consequences.

Thorndike's law of effect

Thorndike's law of effect was, however, also a theory of how reward and punishment modify behaviour. This theory, which states that behaviour normally is modified by changing the strength of stimulus–response connections, finds less general acceptance today. A simple experiment suggests one reason for this. A rat is trained to press a lever in a Skinner box, being rewarded with a small quantity of sucrose solution for each press of the lever. Once the response has been established, the rat is removed from the Skinner box. The next day, while in its home cage, the animal is given sucrose solution to drink and shortly thereafter is made ill by an injection of lithium. Once this treatment has established a strong aversion to the sucrose, the rat is returned to the Skinner box, where, despite the opportunity to do so, the animal does not press the lever again. The result is hardly surprising: there is no reason to expect the rat to perform a response whose sole consequence is the delivery of the now aversive sucrose solution. But this behaviour cannot be explained by Thorndike's theory, for according to Thorndike all that the rat learned in the first stage of the experiment was a new stimulus–response habit; stimuli from the Skinner box should, by Thorndike's reasoning, now elicit the response of pressing the lever. Thorndike's stimulus–response theory credits the rat with no acquired knowledge of the connection between pressing the lever and obtaining sucrose; the function of sucrose is merely to strengthen the stimulus–response connection.

That responses are modified by their consequences, therefore, need not call for Thorndike's theoretical account of this fact. It is probably more reasonable to suppose that animals learn about the relationship between their actions and consequences (just as they can also learn about the relationship between any other classes of events), and that they then modify their actions in accordance with the current value of these consequences. The next question to consider is whether this is an entirely general principle of performance, or whether it applies only to some classes of response in some kinds of situations. Why, for example, does Pavlov's dog start salivating to the ticking of the metronome? Is it because the response of salivating is followed by a rewarding consequence? The response is, at first, elicited by the sight of food and is shortly followed by the rewarding consequence of chewing and swallowing the food. But another simple experiment suggests that salivating to the metronome is not strengthened because it is followed by food. The experimenter can turn on the metronome for five seconds on each trial, at the end of which time the dog receives food—but only if it did not start salivating before the arrival of food. Now the response of salivating to the metronome is followed by an undesirable consequence, the cancellation of the food that would otherwise have been delivered on that trial, but the dog still cannot help salivating (at least sometimes) to the metronome. The implication is that salivating is not a response modified by its consequences, but one reflexly elicited by food and also by any stimulus associated with food. Voluntary responses can be modified by their consequences; involuntary responses (such as blushing when a person is embarrassed or the release of adrenalin when a person is angry or afraid) cannot. The reason Pavlov's dog starts salivating to the metronome is, just as Pavlov himself supposed, that the association between metronome and food means that the metronome can substitute for food. To put it another way, the metronome now produces activity in neural centres normally responsive to the delivery of food, activity that is reflexly connected to the salivary response.

It should not be thought that only autonomic, glandular responses are involuntary in this sense. If a small light is always illuminated for five seconds before the delivery of food to a hungry pigeon, the pigeon will learn, by classical conditioning, to approach and peck at the light. Exactly the same experiment as that described above can be undertaken, with food delivered only on those trials when the pigeon does not approach and peck the light during the initial five seconds. The pigeon cannot help doing so. Pavlovian conditioning appears to be a widespread phenomenon, applying to a relatively wide range of responses.

Functions of conditioning. The behaviour of the dog and pigeon in the above experiments seems maladaptive, precisely because it violates the law of effect. If the way to obtain food is to refrain from performing a particular response, then that is what the law of effect says the animal should do. The law of effect makes obvious adaptive sense; several writers, indeed, have pointed to the analogy between the law of effect and natural selection. Just as natural selection favours those variations that happen to increase fitness, so the law of effect selects those responses that happen to be followed by certain consequences.

The fact that Pavlovian conditioning may result in apparently maladaptive behaviour in the artificial confines of the experimental psychologist's laboratory, however, does not mean that it is not adaptive in the real world. The pigeon's behaviour provides a clue. In a normal classical conditioning experiment, where the illumination of a small light regularly precedes the delivery of food, the pigeon will rapidly learn to approach and direct pecks at the light. Approach and pecking are food-related activities: what is happening is that a simple process of Pavlovian conditioning is ensuring that responses related to food are being elicited by stimuli associated with food. It is not difficult to appreciate the adaptive significance of a process that results in animals approaching places where they have found food in the past, or in learning that a particular novel object is in fact an example of food, and directing food-related activity toward these stimuli in the future.

Adaptive value of Pavlovian conditioning

Pavlovian conditioning also affects other significant behaviours. For example, it probably provides the basic process by which animals learn to avoid poisonous foods. If a novel food is associated with illness, its taste will elicit responses of disgust or nausea, ensuring that the substance will subsequently be rejected after the first taste.

In territorial birds and fish, aggressive displays and attacks can become conditioned to stimuli that regularly precede the appearance of a rival male. A male already primed to threaten and attack an intruder, because he has learned that certain signs herald the appearance of the intruder, should be more successful in defense of his territory than the male that is unprepared. Experimental analysis has, in fact, nicely confirmed this expectation. In general, any pattern of defensive behaviour that is adaptive in response to an intruder or predator—such as displaying or fighting, fleeing or taking other evasive action, or freezing into immobility or feigning death—will be even more adaptive if performed in advance, at the first reliable signal of the predator's or intruder's appearance.

The process of Pavlovian conditioning thus often enables animals to behave appropriately in anticipation of events of biological significance, without involving any direct modification of that behaviour by its success or failure.

Adaptive value of instrumental conditioning

But further modification must sometimes be of further advantage. For instance, it is not always enough just to approach a stimulus associated with food; if that stimulus is a prey species, it may take evasive action that will require much more elaborate behaviour on the part of the predator. This can be seen in the feeding behaviour of the oystercatchers, a group of birds that eat bivalve mollusks. Oystercatchers first catch their pray by probing down the hole made by the bivalve in the mud; the sight of the hole must be rapidly established as a conditional stimulus for food. But the birds must then perform a complex series of actions to get at the mollusk's flesh, and this skilled sequence of responses also must be learned, presumably in accordance with the law of effect. Similarly, many animals have a wide range of defensive behaviour patterns; in the laboratory, at least, which one eventually predominates in any given situation normally depends on which one successfully enables the animal to escape or to avoid aversive consequences. In all these cases, it appears that instrumental conditioning serves to modify, via the law of effect, initial responses that owed their origin to Pavlovian conditioning.

The adaptive value of instrumental conditioning is an area of research that has seen some fruitful collaboration among experimental psychologists, ethologists, and behavioral ecologists. From ecology has come the "optimal foraging theory," the idea that efficient foraging behaviour should maximize an animal's net rate of food intake. From ethology and experimental psychology has come the idea that an animal's instrumental behaviour in any given situation is a product of competition between various possible activities, a competition whose resolution depends on weighing the costs and benefits of increasing one activity at the expense of another. Both in the laboratory and in more natural settings, for example, the proportion of time spent searching for one kind of food depends not only on the probability of finding that food and on its value when found but also on the probability of the animal finding an alternative food if it looks elsewhere. There is also abundant evidence that animals improve their foraging efficiency with practice; this clearly must depend on learning which stimuli signal the availability of which kinds of food, the most efficient way of taking a given food, and the most effective distribution of time between alternatives.

SPATIAL LEARNING

One of the major problems many animals must confront is how to find their way around their world—for example, to know where a particular resource is and how to get to it from their present location, or what is a safe route home to avoid a predator. Such spatial learning may cover only the highly restricted confines of an animal's home range or territory, or it may embrace a migration route of several hundreds or even thousands of miles. Although some forms of navigational behaviour may be explicable in relatively simple terms, not necessarily requiring appeal to processes more complex than those of simple conditioning, others suggest some quite new principles.

Maze learning. In the psychologist's laboratory, the primary method of studying spatial learning has been to put a rat in a maze and watch how it finds its way to the goal box, where it is fed. As befits the analytic (some would say sterile) approach so popular in experimental psychology, the elaborate and complex mazes used in earlier studies (the very first published experiment used a scaled-down replica of the maze at Hampton Court, London) soon gave way to something very much simpler, a T-maze or Y-maze. A rat placed at the end of one arm must run to the central choice-point, from where it has to enter one of the two remaining arms. Although extremely simple, even this apparatus allows for a number of possible modes of solution. One possibility is that the rat learns to execute a particular response, a left turn or a right turn, at the choice-point, because that response is followed by food. A second possible solution is that the rat learns that the two alternative arms differ in some particular way and further learns to associate one of the arms with food and hence to choose it. The third and most interesting possibility is that the rat learns to define the rewarded arm not in terms of its own intrinsic characteristics but by its spatial relationship to an array of landmarks outside the maze. Thus the rat might learn that the correct arm is the one pointing to the left of a window and away from a table with a lamp on it. Experiments show that whenever such landmarks are available, this third solution mode is the one used.

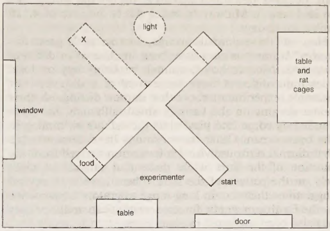


Figure 3: Diagram of a T-maze. Once a rat has learned to run to the arm that contains food, the maze can be flipped (so that the start-arm is in the position marked X) to determine if the rat has learned to reach the goal by making a left turn. If the rat turns right on this trial, the experimenter can rotate the maze 180° to see if the animal has learned to identify the goal-arm in terms of its intrinsic characteristics or, as is usually the case, by its relation to landmarks outside the maze.

Evidence of spatial learning in rats

Perhaps the most convincing demonstration that rats can find their way to a particular location—one defined solely in terms of its spatial relation to various external landmarks—has been provided by experiments in which the animals are placed in a large circular tank of water and must swim to a transparent platform submerged somewhere in the middle of the tank. They can rapidly learn to do this, regardless of where they are initially put into the tank and even though the platform itself is invisible. (The invisibility of the platform is shown by the following: if the platform is moved, the rat will swim straight past it, heading instead toward the position it used to occupy.)

Rats in these experiments are not simply approaching a single landmark; they locate their goal by reference to its spatial relationship with a whole series of landmarks, no one of which is necessary. This can be established by using half a dozen arbitrary but easily identified objects as landmarks during maze training. Removal of any one or two of them in no way disrupts the rat's behaviour. If all the landmarks are systematically rotated around the room, the rat will identify a new arm of the maze as correct (the one that has the same relationship to the landmarks as the initially correct arm). If, however, the landmarks are rearranged in such a way as to destroy their original spatial relationship to one another, the rat does not know which arm to choose.

The processes involved in this sort of learning are not well understood. Some psychologists have been sufficiently impressed by the rat's flexibility in these experiments to argue that the animal is constructing a map of its environment—not, obviously, a written map but an internal, maplike representation that encodes a complete set of spatial relationships between major landmarks. The best evidence for such a maplike representation would be if a rat could take an unfamiliar route when its original route to a goal is blocked. Unfortunately, there is little evidence of such performance in rats, except in the not especially critical case where the goal, or a stimulus very close to it, is clearly visible from the choice-point. On the other hand, studies of long-range navigation have shown that some animals can do just this.

Navigation. Salmon return from the ocean to spawn in the stream in which they were hatched; swallows return to the same nest sites in northern Europe each spring from wintering in southern Africa. These and other examples of large-scale migrations have long fascinated students of animal behaviour, and experimental intervention has produced some remarkable results. A Manx shearwater was taken in an airplane from its breeding site on the island of Skokholm, off south Wales, to Boston, Mass. It returned to Skokholm within 13 days of being released in Boston; the direct distance between these two points is 3,050 miles, which implies (assuming that the bird did not fly at night) a minimum average speed in excess of 20 miles per hour. An albatross flew from a release site in the Philippines to its home in Midway Island, a direct distance of 4,120 miles, in 32 days.

Navigational cues

How do these animals navigate across such great distances? Numerous cues have been implicated in different instances. Near to home, animals probably rely on local cues quite different from those used at a distance. For example, experiments show that salmon distinguish their home streams on the basis of smell, although this sense can hardly come into play while the fish are swimming in the open ocean. Other investigations have demonstrated that diurnal birds use visual information derived from the position of the Sun, while those that migrate at night rely on the pattern of the stars. There have been several suggestions that certain long-range migrators are sensitive to the Earth's magnetic forces; sensitivity to auditory cues has also been suggested in some cases.

The most intensive analysis of long-range navigation has been undertaken with homing pigeons. These birds are trained by being released from sites progressively further from their home loft. Just what the pigeons learn on these training flights is not entirely clear. In part, they obviously learn the visual landmarks immediately surrounding the home loft, but experimental evidence suggests that they use such landmarks only very close to home. Once some training has been given, however, a pigeon can be taken 100 miles or more in any direction from home, and it will, within a few minutes of its release, start flying in a homeward direction.

One general class of theory on homing behaviour postulates that the pigeon detects a discrepancy between a particular set of stimuli observed at the release site and its stored knowledge of what that set of stimuli should be like at home, and it then flies in such a direction as to reduce this discrepancy. Different versions of this theory appeal to different sets of stimuli that might be used to guide the pigeon home. At one time, a popular idea was that the pigeon used the Sun's height in the sky in combination with an internal clock. At any given season and time of day, the Sun's height in the sky—and, by extrapolation from its current rate of climb, its maximum height—are unique to a single place (in this case, the pigeon's home). Assuming that the pigeon's home loft and the release site are both in the Northern Hemisphere, then if the Sun's maximum height is lower at the release site, the release site is north of home; if higher, then the release site is south of home. If the Sun will reach its maximum height later than at home, the release site is west of home; if earlier, the site is east of home. If released at noon at a site in the Northern Hemisphere 200 miles northeast of home, the pigeon must fly so as to raise the maximum height of the Sun (*i.e.*, south), and so as to stop the Sun falling (*i.e.*, west).

This explanation is immensely ingenious and, although calling for some astonishingly fine sensory discriminations on the part of the pigeon, not impossible in principle. Unfortunately, it is probably wrong. Two critical experiments have produced results quite at variance with its predictions. The first suggested that pigeons do not rely on the height of the Sun to navigate at all. In this experiment, the pigeons were confined to a laboratory from which they could see the Sun for only a relatively short time around noon each day, and the apparent height of the Sun above the horizon was raised or lowered by allowing the birds to view the Sun only through a complex series of mirrors. This should have had drastic effects on their perception of the true position of their home; for instance, an increase of 70′ in the apparent height of the Sun at noon would correspond to an 80-mile southward relocation of the home. The pigeons were then taken from home and released 40 miles south, where they saw the real Sun for the first time in several weeks. If the Sun's height was indeed a critical stimulus in navigation, the pigeons would be expected to fly south rather than north. In fact, they correctly flew north.

Experiments on navigation in homing pigeons

The second experiment involved shifting the birds' internal clock, by confining them indoors and exposing them to a new light–dark cycle. Independent observations had shown that this procedure is entirely successful: if a bird is confined indoors for a few weeks with the lights switched on every day at midnight and switched off at noon, its clock soon will be entrained on this new cycle, so that 6:00 AM is regarded as the middle of the day. In the critical experiment, the bird was taken out of the laboratory and released at 6:00 AM (true time) from a site 50 miles south of home. The Sun, now seen for the first time in several weeks, was just rising; but, according to the pigeon's internal clock, the time at home was noon. This implied that the release site was a long way west of home, and if the pigeon were using the height of the Sun as a cue to guide it home, it should have flown east. In fact, the pigeon flew west.

The result of the second experiment indicates that the pigeon was using the position of the Sun in the sky, and that the clock shift had been effective (for the pigeon was not flying in the direction of home). This is readily explained by the hypothesis that the pigeon used the Sun as a compass. If we allow, for the sake of argument, that the pigeon knew that the release site was south of home, then it should have tried to fly north. In the Northern Hemisphere at noon, the Sun is due south; therefore, the pigeon—whose internal clock said it was noon—should fly away from the Sun. But although the pigeon's shifted clock said that it was noon, the true local time was 6:00 AM, and the Sun was in the east. Flying away from the Sun, the pigeon flew west. This experiment then suggests first, that the pigeon was not using the height of the Sun at all; second, that it used the Sun's horizontal position, or azimuth, to provide a compass bearing; and, third, and most important, that the pigeon had some other map that told it that the release site was south of home. In general, a compass is of no use without a map.

Use of the Sun as a compass

The basis for the map component of the pigeon's navigational skill remains extremely obscure. There is evidence from studies of many migratory birds that the compass component is in some sense innate, but that a map of the relative positions of the summer and winter habitats and of other places in between (or even not in between) develops only with the experience of migration. For example, starlings that breed around the Baltic Sea fly southwest in autumn to winter in southern England, northern France, and Belgium. When captured during this autumn migration and released in Switzerland (some 500 miles south of their normal route), experienced, adult birds flew back to northern France and Belgium—even though they had presumably never flown over any part of this route before. Young birds, however, for whom this was the first migration, flew southwest from Switzerland and ended up in southern France or northern Spain. They clearly had a compass that told them which direction was southwest; what they lacked was any knowledge of the spatial relationship between their present location in Switzerland and their goal in northern France.

The role of experience in establishing an internal map

PERCEPTUAL LEARNING

According to Thorndike's stimulus–response theory, learning, which is reducible to the strengthening and weakening of the tendency to perform a particular response in the presence of a particular stimulus, occurs only when that response is performed; learning, in other words, depends on trial and error. Even in the realm of simple conditioning, there are good reasons to question this restriction. Conditioning is better conceptualized as the acquisition of knowledge about temporal relationships between events rather than as the acquisition of behaviour. Spatial learning seems to be a matter of learning about spatial relationships between objects and places in one's environment and, apparently, the construction of some sort of map that will subsequently permit the animal to perform a new sequence of actions across unknown territory. This section considers other examples of learning, in which at least part of what an animal appears to acquire is the recognition of a more or less complex set of stimuli that subsequently can be used to guide its actions.

Imitation and observational learning. One reason why Thorndike adopted such a narrow, behavioral view of learning was that he looked for evidence of other forms of learning without success. Having taught one cat to escape from the puzzle box by operating a latch, he looked to see whether a second cat would acquire the correct solution simply by watching the first. A series of such experiments produced uniformly negative results, and Thorndike concluded that trial and error was the only form of learning available to animals other than humans.

Why Thorndike should have been so unsuccessful is something of a mystery, for later experiments have established quite convincingly that animals can often benefit from watching another member of their species perform a particular task. Casual observation in natural settings, for instance, reveals that young chimpanzees intently watch their elders perform intricate tasks; this certainly suggests that learning by observation is very common in some species.

Experimental analysis has revealed a number of important distinctions concerning the role of observation in behaviour. For example, domestic chickens that have eaten to satiation a particular source of food will start eating again if they observe other chickens feeding. Although the observation of conspecifics engaged in a particular activity has clearly affected the tendency of the satiated chicken to engage in that activity, it is not clear what they might have learned from this observation. They already know how to peck, and they already know that the grain before them is palatable food. It is probably more appropriate to regard this as an instance of "social facilitation" and to say that one of the stimuli that elicits feeding in chickens is the sight of other chickens feeding.

The basic nature of observational learning

The example above demonstrates the minimum requirement for establishing that an animal has learned by observation: in the absence of the opportunity to observe another, the animal must have been unlikely to have performed a particular response, and the reason for this must reside in lack of knowledge. An artificial, laboratory example of observational learning would be to allow an observer rat to watch a demonstrator rat pressing a lever for food. If the observer has never before pressed a lever and, given the opportunity, now does so much more rapidly than another rat denied the opportunity to observe the demonstrator, surely some genuine observational learning has occurred. But even here it remains difficult to establish exactly what it is that the observer has learned by watching the demonstrator, and more elaborate experiments may be required to elucidate this. An experiment with two monkeys showed how this may be done. The monkeys took turns acting as demonstrator and observer. The demonstrator's task was to choose between two objects, one of which contained some hidden food. Since the objects were changed on each new trial for the demonstrator, there was no way for the animal to know which choice was correct, and it necessarily picked one at random. The observer, however, could watch the demonstrator's trial and thus could find out which of the two objects in a particular set was correct. Given an opportunity to choose between the two, the observer more often than not chose correctly. That the observer was not simply watching the demonstrator, but was in fact looking to see the outcome of the choice, is established by the finding that the observer performed somewhat more accurately on those trials when the demonstrator's choice was wrong than on those when it was right.

The distinction between observation and imitation

This last finding points to a further distinction, that between observing the actions of another and imitating those actions. In this particular experiment, the monkeys clearly were not imitating one another, or they would have copied each other's choices even when these were wrong. A demonstration of imitation is provided by the behaviour of oystercatchers feeding on mussels. Having found a mussel, an adult oystercatcher obtains the food from within either by inserting its beak in the right place and cutting the muscle that holds the shell together or by pecking a hole in the weakest point of the shell. Young birds develop the method employed by their parents, but experiments in which chicks were fostered by adults with a different habit from that of the natural parents have established that this behaviour is not genetically determined. Rather, the young birds imitate the actions they observe being performed by their foster parents.

The best known natural example of such imitation was provided by a troop of macaques in Japan. In order to lure the monkeys out of the forest and into the open, where their behaviour could be better studied, scientists routinely left sweet potatoes and wheat on the beach. The monkeys ate this food but clearly disliked the fact that it had become liberally mixed with sand. A young female member of the troop, however, discovered that sweet potatoes could readily be washed free of sand, and that a handful of wheat and sand could be thrown into a pool, where the sand would sink, leaving the wheat floating behind. Both customs spread through the troop, first to the immediate family and young companions of the original inventor, and last of all (an interesting touch) to the old, conservative males. Other examples of observational learning are readily apparent in the behaviour of animals in the field, but in many cases, as in some of the laboratory studies cited above, it remains difficult to elucidate just what it is that has been learned.

Song learning. A special case of observational learning is that of young birds acquiring their species-typical song. Numerous species of animals, including many birds, produce species-typical calls or other vocalizations as adults; in many cases, however, there is little evidence that learning plays any significant role in their development. In many species of crickets, for example, the song is stereotyped, and the pattern of neural activity that produces the song can be detected even in young animals who neither sing nor apparently react to the adult song. But in most songbirds, there is reason to believe that learning has a significant effect on the development of the adult song.

The interesting feature of this learning is that it sometimes occurs in two distinct phases separated by several months. The first of these can be regarded as purely observational learning, the second as the perfection of the song through practice (*i.e.*, as imitation of a model). Song sparrows, for example, do not develop a normal adult song unless they have the opportunity to hear the song during their first autumn. There is thus a sensitive period during which they must hear their species' song if they are to develop normally, but it is important to note that they do not themselves sing at all during the first autumn. It is not until the next spring that they start practicing the song. At this point, they do not need to hear other sparrows singing, but they do need to hear themselves. If the bird is deafened before it starts practicing, only a very crude song emerges. The implication is that, during exposure in the first autumn, the sparrow learns to identify the detailed song and establishes a template of it; the following spring, the sparrow starts singing and needs practice to match its output to the stored template.

The role of observation

The song sparrow provides an example of a particularly clear separation between observation and imitation. In other species, such as the chaffinch, the young bird learns from exposure to song in the first autumn, but refinement of the song is produced by further exposure to other chaffinches singing during the following spring. In yet others, such as indigo buntings, the adult bird learns its song from territorial neighbours. But even where there is no temporal separation between the two aspects of learning, it still seems valid to distinguish between the learning involved in establishing the template and that involved in perfecting the motor skill.

If song learning consists solely of the young bird learning to reproduce the adult, species-typical song, one might wonder why any learning should be necessary at all. Why should the song not develop simply through maturation, or, in other words, why is not the template, at least, genetically laid down in the bird's brain? In fact, studies indicate that a relatively crude template is innately determined in most species. There are very strict limits to the range of songs that a bird of one species can learn. Moreover, among chaffinches and certain other species, even if a young bird hears no song at all it will still develop a crude song that has recognizable features of the full, species-typical one. The degree of this innate specification varies widely from species to species: at one extreme are such birds as cuckoos, which develop a standard call with no prior exposure at all; at the other extreme are such birds as marsh warblers, which develop idiosyncratic songs picked up, it seems, from any other species they come in contact with during the sensitive period.

Value of learned songs

Species whose song acquisition involves a great deal of individual learning are probably those in which individual birds develop slightly different songs. In some species, such as song sparrows, there are recognizable local "dialects" that the young birds learn from adults living in the same region. In other species, there is even more variation between individuals. If one function of the song is to attract a mate, then an interplay is called for between a song that simply advertises the singer's species and one that establishes his individual identity. The importance of individual learning, then, depends on the role of the song in the mating patterns of the species.

Imprinting. The young of many species are born relatively helpless: in songbirds, rats, cats, dogs, and primates, the hatchling or newborn infant is wholly dependent on its parents. These are altricial species. In other species, such as domestic fowl, ducks, geese, ungulates, and guinea pigs, the hatchling or newborn is at a more advanced stage of development. These are precocial species, and their young are capable, among other things, of walking independently within a few minutes or hours of birth, and therefore of wandering away from their parents. Since mammals are dependent on their mothers for nourishment, and even birds are still dependent on parental guidance and protection, it is important that the precocial infant not get lost in this way. The phenomenon of filial imprinting ensures that, in normal circumstances, the precocial infant forms an attachment to its mother and never moves too far away.

Value of imprinting

Although imprinting was first studied by the Englishman Douglas Spalding in the 19th century, Konrad Lorenz is usually, and rightly, credited with having been the first not only to experiment on the phenomenon but also to study its wider implications. Lorenz found that a young duckling or gosling learns to follow the first conspicuous, moving object it sees within the first few days after hatching. In natural circumstances, this object would be the mother bird; but Lorenz discovered that he himself could serve as an adequate substitute, and that a young bird is apparently equally ready to follow a model of another species or a bright red ball. Lorenz also found that such imprinting affected not only the following response of the infant but also many aspects of the young bird's later behaviour, including its sexual preferences as an adult.

Imprinting, like song learning, involves a sensitive period during which the young animal must be exposed to a model, and the learning that occurs at this time may not affect behaviour until some later date. In other words, one can distinguish between a process of perceptual or observational learning, when the young animal is learning to identify the defining characteristics of the other animal or object to which it is exposed, and the way in which this observational learning later affects behaviour. In the case of song learning, observation establishes a template that the bird then learns to match. In the case of imprinting, observation establishes, in Lorenz' phrase, a model of a companion, to which the animal subsequently directs a variety of patterns of social behaviour.

With imprinting, as with song acquisition, one can ask why learning should be necessary at all. Would it not be safer to ensure that the young chick or lamb innately recognized its mother? There are, in fact, genetic constraints on the range of stimuli to which most precocial animals will imprint. A model of a Burmese jungle fowl (the species whose domestication produced domestic chickens) serves as a more effective imprinting object for a young chick than does a red ball; there is even evidence that imprinting in the latter case involves different neural circuits from those involved in imprinting to more natural stimuli. Nonetheless, it is clear that the innate constraints are not very tight and that a great deal of learning normally occurs. The most plausible explanation, as in the case of song learning, is that imprinting involves some measure of individual identification. Lorenz argued that one of the unique characteristics of imprinting was that it involved learning the characteristics of an entire species. It is true that imprinting results in the animal directing its social and mating behaviour toward other members of its own species, and not necessarily toward the particular individuals to which it was exposed when imprinting occurred. But learning usually involves some generalization to other instances, and there does not seem to be anything peculiar to imprinting here. The primary function of imprinting, however, is to enable the young animal to recognize its own mother from among the other adults of its species. This no doubt is particularly important in the case of such animals as sheep, which live in large flocks. Only learning could produce this result.

There is also an important element of individual recognition in at least some cases of imprinting's effects on sexual behaviour. Experiments with Japanese quail have shown that their sexual preferences as adults are influenced by the precise individuals to whom they are exposed at an earlier age. Their preferred mate is one like, but not too like, the individuals on whom they imprinted. The preference for some similarity presumably ensures that they attempt to mate with members of their own species. The preference for some difference is almost certainly a mechanism for reducing inbreeding, since young birds will normally imprint on their own immediate relatives.

The possible role of rewards in imprinting

The difference between imprinting and song learning lies in the consequences of observational learning. The effect of imprinting is the formation of various forms of social attachment. But what mechanism causes the young chick or duckling to follow its mother? Lorenz thought that imprinting was unrewarded, yet the tendency of a young bird to follow an object on which it has been imprinted in the laboratory can be enhanced by rewarding the bird with food. Rewards also occur outside the laboratory: the mother hen not only scratches up food for her young chicks, she also provides a source of warmth and comfort. Moreover, following is also rewarded by a reduction in anxiety. As chicks develop over the first few days of life, they show increasing fear of unfamiliar objects; they allay this anxiety by avoiding novel objects and approaching a familiar one. This latter object must be one to which they have already been exposed—in other words, one on which they have imprinted. Imprinting works because newly hatched birds do not show any fear of unfamiliar objects, perhaps because something can be unfamiliar only by contrast with something else that is familiar. On the contrary, the newly hatched birds are attracted toward salient objects, particularly ones that move. Once, however, a particular object has been established as familiar and its features identified, different objects will be discriminated from it. These will be perceived as relatively unfamiliar, and hence they will provoke anxiety and the attempt to get as close as possible to the more familiar object. The imprinting of the young bird on one object necessarily closes down the possibility of its imprinting on others, as these will always be relatively less familiar. Thus, there is normally a relatively restricted period in the first few hours or days of life during which imprinting can occur. The only way to prolong this period is to confine the newly hatched bird to a dark box where it is exposed to no stimuli; prevented from imprinting during this period of confinement, the bird imprints on the first salient object it sees after emerging.

COMPLEX PROBLEM SOLVING

Experimental psychologists who study conditioning are the intellectual heirs of the traditional associationist philosophers. Both believe that the complexity of the human or animal mind is more apparent than real—that complex ideas are built from simple ideas by associating simple elements into apparently more complex wholes. According to this perspective, the only relationship between these ideas is their association, and the determinants of these associations are themselves relatively simple and few in number. Neither conditioning theorists nor associationist philosophers, however, have lacked for critics, who claim that intelligent problem solving cannot be reduced to mere association. Although allowing that the behaviour of invertebrates, and perhaps that of birds and fish, may be understood in terms of instincts and simple forms of nonassociative and associative learning, these critics maintain that the human mind is an altogether more subtle affair, and that the behaviour of animals more closely related to man—notably apes and monkeys, and perhaps other mammals as well—will share more features in common with human behaviour than with that of earthworms, insects, and mollusks.

Differing views of animal intelligence

The idea that animals might differ in intelligence, with those more closely related to humans sharing more of their intellectual abilities, is commonly traced back to Charles Darwin. This is because the acceptance of Darwin's theory of evolution was at the expense of the ideas of the French philosopher René Descartes, who held that there is a rigid distinction between man, who has a soul and can think and speak rationally, and all other animals, who are mere automatons. The Cartesian view had, in fact, been challenged long before Darwin's time by those who believed (as seems obvious from even the most casual observations) that some animals are notably more complicated than others, in ways that probably include differences in behaviour and intelligence. It was, however, the publication of Darwin's *Descent of Man* (1871) that stimulated scientific interest in the question of mental continuity between man and other animals. Darwin's young colleague, George Romanes, compiled a systematic collection of stories and anecdotes about the behaviour of animals, upon which he built an elaborate theory of the evolution of intelligence. It was largely in reaction to this anecdotal tradition, with its uncritical acceptance of tales of astounding feats by pet cats and dogs, that Thorndike undertook his studies of learning under relatively well-controlled laboratory conditions. Thorndike's own conclusions, already noted above, were distinctly Cartesian: animals ranging from chickens to monkeys all learned in essentially the same way, by trial and error or simple instrumental conditioning. Unlike man, none could reason.

This controversy actually involves two questions, which are worth keeping apart. The first is whether theories of learning based on the results of, say, simple conditioning experiments are sufficient to explain all forms of learning and problem solving in animals. The second question is whether new and more complex processes operate only in some animals, that is to say, whether some animals are more intelligent than others. The distinction between these questions is not always easy to preserve, for they are clearly related, and an answer to one usually has implications for the other. The remainder of this article is organized around the first question; in cases where the behaviour of an animal does, in fact, seem to indicate that more complex processes are involved, the second question is also considered.

Discrimination of relational and abstract stimuli. Laboratory studies of habituation and conditioning usually employ very simple stimuli, such as lights, buzzers, and ticking metronomes in Pavlov's experiments. Some of the other examples of learning considered earlier have already suggested that animals can actually respond to additional, more complex stimuli. Even the solution of simple spatial discriminations in the laboratory requires the animal to learn about spatial relationships between different landmarks; migration or navigation over hundreds of miles demands abilities at least as complex as this. Song learning requires the young bird to discriminate between different sequences of subtly varying notes and calls, and the individual recognition involved in imprinting requires response to elaborate configurations of features.

Thus, one way in which a problem may become more difficult is if its solution depends on response to more subtle changes in stimuli. Numerous laboratory studies have examined the abilities of a variety of animals to perform such discriminations. The phenomenon of transposition, first studied in chicks by the Gestalt psychologist Wolfgang Köhler, suggests that animals may solve even simple discriminations in ways more complex than the experimenter had imagined. Köhler trained his chicks to perform simple discriminations—say, to choose a large white circle (five centimetres in diameter) in preference to a small white circle (three centimetres in diameter). He then sought to discover whether the animal was responding to the relationship between the two stimuli or to the absolute characteristics of the stimuli. In other words, had the chick learned to select the larger of the two circles, or had it learned to pick the five-centimetre circle? If the former were the case, Köhler reasoned that given the choice between the five-centimetre circle and an even larger one (eight centimetres in diameter), the animal should transpose the relationship and choose the larger circle. This was indeed the result, demonstrating that the animal was responding in terms of the relationship between stimuli rather than, or at least in addition to, their absolute properties.

Transposition of relational discriminations

Transposition experiments show that animals can respond to relationships between stimuli varying along a particular continuum of physical characteristics: size, brightness, hue, etc. Another question is whether animals can respond to an abstract property of a stimulus array, independent of the actual physical stimuli making up that array. In experiments on counting, the animal must choose between an array containing, say, five stimuli and one containing three. The actual stimuli in the array vary from trial to trial, in order to rule out the possibility that the animal is responding in terms of other features, such as differences in total area or brightness, between the arrays. Counting experiments have been tried on birds more frequently than on any other class of animal, and several species, notably ravens, rooks, and jackdaws, have solved this type of problem. This success may not be entirely by chance, for there is reason to believe that the stimulus that controls when a female bird stops laying eggs is something to do with the number of eggs already laid and in the nest. Chimpanzees, however, have been trained to label pictures of various objects (*e.g.*, spoons, shoes, padlocks, and balls) with the numeral specifying the number of objects in the picture. Moreover, rats and other standard laboratory animals have solved similarly abstract discriminations, for example, of temporal duration. A rat can learn to perform one response after a stimulus has been turned on for two seconds and a different response after the stimulus has been turned on for five seconds. The nature of the actual stimuli employed can vary without disrupting the rat's discrimination, suggesting that it is the duration of the stimuli to which the rat responds.

Counting as an example of an abstract discrimination

Concept learning

Concept learning makes up another class of discriminations that may be solved by the abstraction of a particular property or set of properties from a very wide array of individual stimuli. In a typical experiment, a pigeon is shown a large number of colour photographs of natural scenes: half of these contain, somewhere within the scene, all or part of a tree or group of trees; the other half contain no tree (although there might be flowers, a climbing rose, or other plants). Responding to the pictures of trees is rewarded, but responding to the remaining pictures is not. Pigeons rapidly learn the discrimination. In one sense, perhaps this is not surprising: birds that roost in trees, one is inclined to argue, must be able to recognize them. But pigeons can learn other discriminations with almost equal facility; for example, they can be trained to distinguish between underwater scenes containing a fish and similar views with no fish present. In such cases, the class of stimuli in question is one for which their evolutionary history can hardly have prepared pigeons. The question, of course, is how the pigeons solve such problems. Are they, in some sense, abstracting a conceptual rule for categorizing the world into classes of stimuli? Or are they responding to what is no doubt a very large number of particular features that differentiate trees or fish from other objects in the world?

Pigeons, in common with most birds, rely more heavily on vision, and certainly have better developed colour vision, than most mammals—with the exception of primates. There is evidence that monkeys can solve the concept discriminations that have been set to pigeons, but there is no evidence that other mammals can. For extensive comparative analysis, therefore, it is necessary to turn to different kinds of tasks. One that has been studied almost to excess is discrimination reversal. In reversal tasks, an animal is first trained on a simple discriminative problem: for example, to choose the left-hand arm of a T-maze, where it is rewarded, rather than the right arm, where it is not. Once the animal has solved the problem, the experimenter reverses the reward assignments, so that the food is now in the right arm rather than the left. Training continues until the animal has learned this reversal, whereupon the assignment of reward is switched back to the left arm. And so on. Rats trained on this series of reversals eventually become extremely adept at the task. Although the initial reversal causes considerable problems, with animals making many more errors than on the original discrimination, after a few more reversals these difficulties vanish. Eventually, rats solve each new reversal in fewer trials than they took to solve the original discrimination, often with no more than a single error.

The learning of discrimination reversals

Similarly efficient performance has been observed in a relatively wide range of mammals. More interesting was the early suggestion that the few species of fish (goldfish, African mouthbreeders, and Paradise fish) trained on similar problems showed no evidence of the increase in efficiency displayed by mammals. The fish would learn the first reversal slowly and laboriously, and the 20th reversal equally slowly. Subsequent experiments have established that this was an unfairly pessimistic assessment,

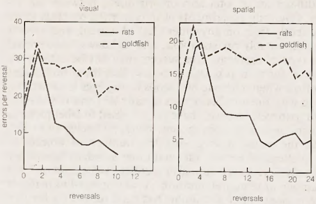


Figure 4: Learning of visual and spatial reversals by rats and goldfish. Although disrupted by the first few reversals of a discrimination, rats soon adjust and learn to reverse their choices with only a few errors. Goldfish also show some adjustment to repeated reversals (they learn later reversals significantly faster than earlier ones), but they do not seem to become as proficient as rats.

for improvements in experimental techniques have been accompanied by a significant improvement in the fish's performance, a finding that highlights the extreme difficulty of assessing the relative efficiency of widely differing animals on supposedly the same task. Nevertheless, it remains doubtful that goldfish are as adept at reversal tasks as rats are.

The theoretical question, however, is how rats attain such efficiency. What processes allow them eventually to learn the reversal of a discrimination faster than they originally learned the discrimination itself, and often with only a single error? The most plausible suggestion is that they develop a "win–stay, lose–shift" strategy. They learn, in other words, to characterize the alternatives between which they must choose not in terms of their physical features but in terms of whether or not they chose it on the previous trial. They then learn that, if the alternative they chose on the last trial was rewarded, choice of that alternative will be rewarded again on the current trial; while, if it was not, choice of the other alternative will now be rewarded. A variety of other experiments have shown that rats can rapidly learn to use the outcome of one trial to predict the outcome of the next, and hence keep track of regular sequential dependencies in the availability of food or other rewards.

The theoretical basis of reversal learning

Generalized rule learning. Second only to the reversal task in popularity as a tool for the comparative analysis of learning has been the learning set task. The latter is designed to measure the animal's ability "to learn to learn"—in other words, to discover whether after having learned a new behaviour the animal can then more readily learn other related behaviours. For example, an animal is trained on a simple discrimination between two objects, A and B. Once the problem has been solved, the experimenter substitutes a new pair of objects, C and D, for the original pair; when the animal has solved this new problem, yet another new pair, E and F, is substituted, and so on. Rhesus monkeys trained on such a series of problems become progressively more efficient at solving each new problem. Like rats trained on reversal tasks, the monkeys eventually solve each new problem after a single trial, choosing at random on the first trial with each new pair of stimuli but thereafter selecting with essentially perfect accuracy.

Learning set tasks

Performance on learning sets, as on reversals, was once thought to discriminate between more intelligent and less intelligent animals. Apes and rhesus monkeys were extremely efficient at such tasks, more so even than New World monkeys, who were, in turn, more efficient than any nonprimate mammals. Again, however, there are grave difficulties in the way of making valid comparisons. Primates have better developed visual systems than most other mammals, so it is not surprising that they should be better at solving a series of visual discrimination problems. Even the difference in performance between rhesus and cebus monkeys (Old World versus New World monkeys) turns out to be attributable to differences in colour vision more than anything else. Rats appear to solve learning set tasks very efficiently if olfactory stimuli are used.

Nevertheless, there may be important intellectual differences also underlying the differences in performance. One reason for thinking so arises from consideration of the processes probably involved in mastering learning sets. The win–stay, lose–shift strategy that explains the progressive improvement in reversal learning can also explain the same improvement in the learning set task—but only if the animal can generalize the strategy to novel stimuli. Successful performance requires that the animal learn that the alternative chosen on the last trial, and the outcome of that choice, predict which alternative will be rewarded on this trial, whatever the nature of the alternatives. Some evidence suggests that primates can generalize rules of this sort more readily than many other animals can. Monkeys trained on a series of reversals of a single discrimination will learn the reversal of any new discrimination with equal facility. By contrast, cats trained on comparable problems show little evidence of such transfer.

A discriminative problem widely used in the study of transfer is the "matching-to-sample" discrimination. A pigeon, for example, is required to choose between two disks, one illuminated with red light and the other with green light. The correct alternative on any one trial depends on the value of a sample stimulus, which is also part of each trial. If this third light is red, then the red disk is correct; if green, then green is correct. The correct alternative is the one that matches the sample. Although naturally more difficult than the simple red–green discrimination, matching-to-sample discriminations are learned readily enough by a wide variety of animals; however, there appear to be differences among animals in their capabilities to transfer

Matching-to-sample discriminations

this learning to a new set of stimuli. Primates and dolphins have shown good evidence of such transfer, but pigeons have shown at best only limited transfer. If pigeons are trained with two or three colours to the point where they are responding with essentially no errors, a substitution of a new colour for one of the trained colours may result in a complete breakdown in the discrimination; there is even some question as to whether they can learn a new matching-to-sample discrimination with new stimuli any faster than pigeons with no prior experience of matching problems.

The abilities to respond in terms of certain relationships between stimuli, to abstract those relationships and invariant features from a complex and changing array of stimuli, and, above all perhaps, to transfer such learning to a completely novel set of physical stimuli seem to be some of the more important processes underlying the solution of complex discriminative problems. The fact that certain evidence suggests that animals may differ in some of these abilities has implications for studies of other forms of problem solving.

Insight and reasoning. Köhler's best known contribution to animal psychology arose from his studies of problem solving in a group of captive chimpanzees. Like other Gestalt psychologists, Köhler was strongly opposed to associationist interpretations of psychological phenomena, and he argued that Thorndike's analysis of problem solving in terms of associations between stimuli and responses was wholly inadequate. The task he set his chimpanzees was usually one of obtaining a banana that was hanging from the ceiling of their cage or lying out of reach outside the cage. After much fruitless endeavour, the chimpanzees would apparently give up and sit quietly in a corner, but some minutes later they might jump up and solve the problem in an apparently novel manner—for example, by using a bamboo pole to rake in the banana from outside or, if one pole was not long enough, by fitting one pole into another to form a longer rake. Other chimpanzees reached the banana hanging from the ceiling by using a wooden box, or a series of boxes stacked precariously on top of one another, as a makeshift ladder.

Köhler's examples of insight learning

Köhler believed that his chimpanzees had shown insight into the nature of the problem and the means necessary to solve it. According to Köhler's interpretation, the solution depended on a perceptual reorganization of the chimpanzee's world—seeing a pole as a rake, or a series of boxes as a ladder—rather than on forming any new associations. But subsequent experimental analysis has cast some doubts on Köhler's claims. The critical observation is that the sorts of solutions that Köhler took as evidence of insight quite clearly depend on relevant prior experience. Chimpanzees will not fit two poles together to form a rake or stack boxes up to form a ladder unless they have had a great deal of prior experience with those objects. This experience may well occur during play, when the young chimpanzee discovers that using a stick can extend the reach of an arm, or that standing on a box can put one within reach of high objects. Thus, what Köhler was studying, without knowing it, was probably the transfer of earlier instrumental conditioning to new situations. As we have already seen, the ability to transfer an old solution to a new stimulus situation is an important one, relevant to a wide range of problem-solving activities. This ability is not at all well understood, but it will not necessarily be greatly illuminated by describing it as insight. Certainly it is not a process unique to the great apes: if the component tasks are sufficiently well-structured, even pigeons can put together two independently learned patterns of behaviour to solve a novel problem.

Criticisms of Köhler's interpretation

Combining information from separate sources to reach a new conclusion is one form of reasoning. The paradigm case of reasoning is the solution of syllogisms; for example, when we conclude that Socrates is mortal given the two separate premises that Socrates is a man and that all men are mortal. Employing transitive inference, we can use the premises that Adam is taller than Bertram and that Bertram is taller than Charles to conclude that Adam must be taller than Charles. Reasoning has often been regarded as a uniquely human faculty, one of the few factors, along with the possession of language, that distinguishes us from the rest of the animal kingdom.

But are humans the only animals that can reason? The unsatisfying answer must be that it depends on what is meant by reasoning. In a very general sense, most animals appear perfectly able to arrive at a conclusion based on combining information obtained on two separate occasions. A formal demonstration is provided by an experiment on instrumental conditioning discussed earlier. If rats learn that pressing a lever provides sucrose pellets and later learn that eating sucrose pellets makes them ill, they will subsequently put these two pieces of information together and refrain from pressing the lever. Monkeys and chimpanzees, however, have been trained to solve problems that appear more similar to transitive inference. They are first given discriminative training between pairs of coloured boxes, called, for example, A, B, C, D, E. Confronted with the choice between A and B, they learn that choice of A is rewarded and B is not. When B and C are the alternatives, they learn that B is correct; when C and D are the alternatives, C is correct; and so on. Although choice of A is always rewarded, and that of E never is, the remaining three boxes each are associated equally often with reward and with nonreward. Nonetheless, given a choice between B and D on a test trial, the animals choose B.

Examples of complex reasoning in nonhuman primates

Syllogistic and transitive inference are not the only forms of reasoning: humans also reason inductively or by analogy. Indeed, analogical reasoning problems (black is to white as night is to __?) form a staple ingredient of some IQ tests. One chimpanzee, a mature female called Sarah, was tested by David Premack and his colleagues on a series of analogical reasoning tasks. Sarah previously had been extensively trained in solving matching-to-sample discriminations, to the point where she could use two plastic tokens, one meaning *same*, which she would place between any two objects that were the same, and another meaning *different*, which she would place between two different objects. For her analogical reasoning tasks, Sarah was shown four objects grouped into two pairs, with each pair symmetrically placed on either side of an empty space. If the relationship between the paired objects on the left was the same as the relationship between those on the right, her task was to place the *same* token in the space between the two pairs. Thus in one series of geometrical analogies, a simple problem would display a blue circle and a red circle on the left and a blue triangle and a red triangle on the right; the correct answer, of course, was *same*. But Sarah was equally correct on more complex problems, even when the relationships in question were functional rather than simply perceptual. For example, she correctly answered *same* when the two objects on the left were a tin can and a can opener and the two on the right a padlock and a key.

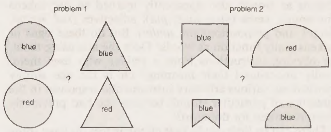


Figure 5: Example of problems in analogical reasoning solved by the chimpanzee Sarah. In each problem her task was to place the token for *same* or *different* between the two pairs of diagrams, depending on whether the relationship between those on the left was the same as, or different from, that between the pair on the right. The correct answer to the first problem is *same;* to the second, *different.*

Solution of analogies requires one to see that the relationship between one pair of items (whether they are words, diagrams, pictures, or objects) is the same as the relationship between a different pair of items. If simple matching-to-sample requires animals to see that one comparison stimulus is the same as the sample and another is different, solving analogies requires them to match relationships between stimuli. The difficulties encountered in training pigeons to generalize simple matching-to-sample discriminations does not encourage one to believe that they would find analogies very easy.

Language learning. The ability to speak was regarded by Descartes as the single most important distinction between humans and other animals, and many modern linguists, most notably Noam Chomsky, have agreed that language is a uniquely human characteristic. Once again, of course, there are problems of definition. Animals of many species undoubtedly communicate with one another. Honeybees communicate the direction and distance of a new source of nectar; a male songbird informs rival males of the location of his territory's boundaries and lets females know of the presence of a territory-owning potential mate; vervet monkeys give different calls to signal to other members of the troop the presence of a snake, a leopard, or a bird of prey. None of these naturally occurring examples of communication, however, contains all of the most salient features of human language. In human language, the relationship between a word and its referent is a purely arbitrary and conventional one, which must be learned by anyone wishing to speak that language; many words, of course, have no obvious referent at all. Moreover, language can be used flexibly and innovatively to talk about situations that have never yet arisen in the speaker's experience—or indeed, about situations that never could arise. Finally, the same words in a different order may mean something quite different, and the rules of syntax that dictate this change of meaning are general ones applying to an indefinite number of other sequences of words in the language.

Defining characteristics of human language

During the first half of the 20th century, several psychologists bravely attempted to teach human language to chimpanzees. They were uniformly unsuccessful, and it is now known that the structure of the ape's vocal tract differs in critical ways from that of a human, thus dooming these attempts to failure. Since then, however, several groups of investigators have employed the idea of teaching a nonvocal language to apes. Some have used a gestural sign language widely used by the deaf to communicate with one another; others have used plastic tokens that stand for words; still others have taught chimpanzees to press symbols on a keyboard. All have had significant success, and several apes have acquired what appears to be a vocabulary of several dozen, and in some cases 100 or 200, "words."

Language learning experiments with apes

Washoe, a female chimpanzee trained by Beatrice and Allan Gardner, learned to use well over 150 signs. Some apparently were used as nouns, standing for people and objects in her daily life, such as the names of her trainers, various kinds of food and drink, clothes, dolls, etc. Others she used as requests, such as *please, hurry,* and *more;* and yet others as verbs, such as *come, go, tickle,* and so on. Sarah, the chimpanzee trained by Premack to use plastic tokens as words, also apparently learned to use tokens for nouns, verbs (*give, take, put*), adjectives (*red, round, large*), and prepositions (*in, under*). But do these signs or tokens really function as words? Does the ape using them, or obeying instructions from a trainer who uses them, really understand their meaning? Or is the ape simply performing various arbitrary instrumental responses in the presence of particular stimuli because she had previously been rewarded for doing so?

There can be little doubt that chimpanzees do have some understanding of what their "words" refer to. Sarah responded appropriately with her token for *red* if asked the question "What colour of apple?" both when an actual red apple was shown as part of the question and when only the token for an apple (which happened to be a blue triangle) was presented. To Sarah, the blue triangle surely stood for, or was associated with, the red apple. In another study, after two chimpanzees had been taught the meaning of a number of symbols for different kinds of food and different tools, they were able not only to fetch the appropriate but absent object when requested to do so, but they could also sort the symbols into two groups, one for foods and one for tools. In another series of studies, a pygmy chimpanzee named Kanzi demonstrated remarkable linguistic abilities. Unlike other apes, he learned to communicate using keyboard symbols without undergoing long training sessions involving food rewards. Even more impressive, he demonstrated an understanding of spoken English words under rigorous testing conditions in which gestural clues from his trainers were eliminated.

As noted above, human language is more than a large number of unrelated words: in accordance with certain implicitly understood syntactic rules, humans combine words to form sentences that communicate a more or less complex meaning to a listener. Can apes understand or use sentences? Undoubtedly they can put together several gestures or tokens in a row. A chimpanzee named Lana, who was trained to press symbols on a keyboard, could type out "Please machine give Lana drink"; Washoe and other chimpanzees trained in gestural sign language frequently produced strings of gestures such as "You me go out," "Roger tickle Washoe," and so on. Skeptical critics, however, have raised doubts about the significance of these strings of signs and symbols. They have pointed out, for example, that when Lana pressed a series of coloured symbols on her keyboard, it was humans who interpreted her actions as the production of a sentence meaning "Please machine give Lana drink." Might it not be equally reasonable to say that she learned to perform an arbitrary sequence of responses in order to obtain a drink? Pigeons can be trained to press four coloured keys—red, white, yellow, and green—in a particular order to obtain food. Psychologists do not feel any temptation to interpret this behaviour as the production of a sentence. What is it about Lana's behaviour that requires this richer interpretation?

Critical interpretations of language experiments

In the case of apes trained to use sign language, two other doubts have been raised. First, there is some reason to believe that a disappointingly high proportion of the apes' gestures may be direct imitations of gestures recently executed by their trainers. Second, a sequence of gestures interpreted as a single sentence is often just as readily interpreted as a number of independent gestures, each prompted, in turn, by a gesture from the trainer. Both these conclusions are based on careful examinations of video recordings of interactions between trainers and apes. Whether they will turn out to be generally true remains an open, and heatedly debated, question.

Without any explicit training, apes have nevertheless learned to produce strings of two or three signs in certain preferred orders: "more drink" or "give me," for example, rather than "drink more" or "me give." Do the animals understand that a string of signs in one order means something different from the same signs in a different order? The following anecdote is suggestive. A chimpanzee called Lucy was accustomed to instructing her trainer, Roger Fouts, by gesturing "Roger tickle Lucy." One day, instead of complying with this request, Fouts signed back "No, Lucy tickle Roger." Although at first nonplussed, after several similar exchanges Lucy eventually did as asked. A simple instance of this sort proves little or nothing, but it may suggest what is needed—namely, that Lucy should understand that changing the order of a set of signs alters their meaning in certain predictable ways. She must generalize the rule that the relationship between the meanings of the signs A-B-C and C-B-A (the same signs in reverse order) is similar to the relationship between the meanings of certain other triplets of signs in her vocabulary when their order is reversed.

The research on language in apes forcefully illustrates a conflict, or tension, that is common to many other areas of research on learning in animals. If the investigators are interested in language and communication, they can attempt to communicate as naturally and informally as possible with their apes. This approach involves treating an ape as a fellow social being, with whom one plays and interacts as far as possible as one would with a human child; it also, almost inevitably, results in a style of research where it is exceptionally difficult to control precisely the cues that the ape may be using and even hard to avoid an overly rich, anthropomorphic interpretation of the ape's behaviour. If, on the other hand, the researchers are interested in rigorous experimental control and economical interpretation of the processes underlying the ape's performances, they are likely to set the ape formal problems to

Conflicts in approach to the study of animal learning solve, with rewards for correct responses and no rewards for errors. But such an approach, however scientific it may seem, must run the risk of missing the point. This is not language; the investigators are not communicating with the ape in the way they would communicate with a child. The very nature of the experimental problems ensures that the ape will not use its language in the way that a child does: to communicate shared interests, to attract a parent's attention to what the child has seen or is doing, to comment on a matter of concern to both.

There is no resolution to this conflict, for both approaches have their virtues as well as their dangers, and both are therefore necessary. In just the same way, the study of a rat pressing a lever in a Skinner box or of a dog salivating to the ticking of a metronome seems to many critics a sterile and narrow approach to animal learning—one that simply misses the point that, if the ability to learn or profit from experience has evolved by natural selection, it must have done so in particular settings or environments because it paid the learner to learn something. It would be foolish to deny this obvious truism: of course it pays animals to learn. Indeed, it may pay them to learn quite particular things in specific situations, and different groups of animals may be particularly adapted to learning rather different things in similar situations. None of this should be forgotten, and the study of such questions requires the scientist to forsake the laboratory for the real world, where animals live and struggle to survive. But few sciences can afford to miss the opportunity to manipulate and experiment under laboratory conditions where this is possible, and none can afford to forget the benefits of precise observation under controlled conditions.

BIBLIOGRAPHY

General works: Historical background is provided by ROBERT BOAKES, *From Darwin to Behaviourism: Psychology and the Minds of Animals* (1984). See also T.R. HALLIDAY and P.J.B. SLATER (eds.), *Animal Behaviour*, vol. 3: *Genes, Development, and Learning* (1983); ROBERT A. HINDE, *Animal Behaviour: A Synthesis of Ethology and Comparative Psychology*, 2nd ed. (1970); J.E.R. STADDON, *Adaptive Behavior and Learning* (1983); and DAVID MCFARLAND, *Animal Behavior: Psychology, Ethology, and Evolution* (1985).

Simple nonassociative learning: GABRIEL HORN and ROBERT A. HINDE (eds.), *Short-Term Changes in Neural Activity and Behaviour* (1970); and HARMAN V.S. PEEKE and MICHAEL J. HERZ (eds.), *Habituation*, 2 vol. (1973).

Associative learning and conditioning: ANTHONY DICKINSON, *Contemporary Animal Learning Theory* (1980); MICHAEL DOMJAN and BARBARA BURKHARD, *The Principles of Learning and Behavior*, 2nd ed. (1986); N.J. MACKINTOSH, *The Psychology of Animal Learning* (1974), and *Conditioning and Associative Learning* (1983); and BARRY SCHWARTZ, *Psychology of Learning and Behavior*, 2nd ed. (1984).

Biological functions of and constraints on learning: ROBERT A. HINDE and J. STEVENSON-HINDE (eds.), *Constraints on Learning: Limitations and Predispositions* (1973); TIMOTHY D. JOHNSTON, "Contrasting Approaches to a Theory of Learning," *Behavioral and Brain Sciences*, 4(1):125–173 (March 1981); and J.R. KREBS and N.B. DAVIES, *An Introduction to Behavioural Ecology* (1981).

Physiological basis of learning: DANIEL L. ALKON and JOSEPH FARLEY (eds.), *Primary Neural Substrates of Learning and Behavioral Change* (1984); and ERIC R. KANDEL, *Cellular Basis of Behavior: An Introduction to Behavioral Neurobiology* (1976).

Spatial learning and navigation: P. ROBIN BAKER, *Bird Navigation: The Solution of a Mystery?* (1984); JOHN O'KEEFE and LYNN NADEL, *The Hippocampus as a Cognitive Map* (1978); and K. SCHMIDT-KOENIG and W.T. KEETON (eds.), *Animal Migration, Navigation, and Homing* (1978).

Song learning and imprinting: P.P.G. BATESON, "The Imprinting of Birds," in S.A. BARNETT (ed.), *Ethology and Development* (1973); HOWARD S. HOFFMAN, "Experimental Analysis of Imprinting and Its Behavioral Effects," *The Psychology of Learning and Motivation*, 12:1–39 (1978); DONALD E. KROODSMA, "Aspects of Learning in the Development of Bird Song," in GORDON M. BURGHARDT and MARC BEKOFF (eds.), *The Development of Behavior: Comparative and Evolutionary Aspects* (1978); and P.J.B. SLATER, "Bird Song Learning: Theme and Variations," in ALAN H. BRUSH and GEORGE A. CLARK, JR. (eds.), *Perspectives in Ornithology* (1983).

Comparative psychology and complex training: E.M. MACPHAIL, *Brain and Intelligence in Vertebrates* (1982); R.E. PASSINGHAM, *The Human Primate* (1982); and H.L. ROITBLAT, T.G. BEVER, and H.S TERRACE (eds.), *Animal Cognition* (1985).

(N.J.M.)

Human Learning and Cognition

A common goal in defining any psychological concept is a statement that corresponds to common usage. Acceptance of that aim, however, entails some peril. It implicitly assumes that common language categorizes in scientifically meaningful ways; that the word *learning,* for example, corresponds to a definite psychological process. However, there appears to be good reason to doubt the validity of this assumption. The phenomena of learning are so varied and diverse that their inclusion in a single category may not be warranted.

Recognizing this danger (and the corollary that no definition of learning is likely to be totally satisfactory) a definition proposed in 1961 by G.A. Kimble may be considered representative: Learning is a relatively permanent change in a behavioral potentiality that occurs as a result of reinforced practice. Although the definition is useful, it still leaves problems.

The definition may be helpful by indicating that the change need not be an improvement; addictions and prejudices are learned as well as high-level skills and useful knowledge.

The phrase *relatively permanent* serves to exclude temporary behavioral changes that may depend on such factors as fatigue, the effects of drugs, or alterations in motives.

The word *potentiality* covers effects that do not appear at once; one might learn about tourniquets by reading a first-aid manual and put the information to use later.

To say that learning occurs as a result of practice excludes the effects of physiological development, aging, and brain damage.

The stipulation that practice must be reinforced serves to distinguish learning from the opposed loss of unreinforced habits. *Reinforcement* objectively refers to any condition—often reward or punishment—that may promote learning.

However, the definition raises difficulties. How permanent is relatively permanent? Suppose one looks up an address, writes it on an envelope, but five minutes later has to look it up again to be sure it is correct. Does this qualify as relatively permanent? While commonly accepted as learning, it seems to violate the definition.

What exactly is the result that occurs with practice? Is it a change in the nervous system? Is it a matter of providing stimuli that can evoke responses they previously would not? Does it mean developing associations, gaining insights, or gaining new perspective?

Such questions serve to distinguish Kimble's descriptive definition from theoretical attempts to define learning by identifying the nature of its underlying process. These may be neurophysiological, perceptual, or associationistic; they begin to delineate theoretical issues and to identify the bases for and manifestations of learning. (The processes of perceptual learning are treated in the article PERCEPTION.)

This article is divided into the following sections:

Theories of learning

THE RANGE OF PHENOMENA CALLED LEARNING

Even the simplest animals display such primitive forms of adaptive activity as habituation, the elimination of practiced responses. For example, a paramecium can learn to escape from a narrow glass tube to get to food. Learning in this case consists of the elimination (habituation) of unnecessary movements. Habituation also has been demonstrated for mammals in which control normally exercised by higher (brain) centres has been impaired by severing the spinal cord. For example, repeated application of electric shock to the paw of a cat so treated leads to habituation of the reflex withdrawal reaction. Whether single-celled animals or cats that function only through the spinal cord are capable of higher forms of learning is a matter of controversy. Sporadic reports that conditioned responses may be possible among such animals have been sharply debated.

At higher evolutionary levels the range of phenomena called learning is more extensive. Many mammalian species display the following varieties of learning.

Classical conditioning. This is the form of learning studied by Ivan Petrovich Pavlov (1849–1936). Some neutral stimulus, such as a bell, is presented just before delivery of some effective stimulus (say, food or acid placed in the mouth of a dog). A response such as salivation, originally evoked only by the effective stimulus, eventually appears when the initially neutral stimulus is presented. The response is said to have become conditioned. Classical conditioning seems easiest to establish for involuntary reactions mediated by the autonomic nervous system.

Conditioning of involuntary reactions

Instrumental conditioning. This indicates learning to obtain reward or to avoid punishment. Laboratory examples of such conditioning among small mammals or birds are common. Rats or pigeons may be taught to press levers for food; they also learn to avoid or terminate electric shock.

Chaining. In the form of learning called chaining the subject is required to make a series of responses in a definite order. For example, a sequence of correct turns in a maze is to be mastered, or a list of words is to be learned in specific sequence.

Acquisition of skill. Within limits, laboratory animals can be taught to regulate the force with which they press a lever or to control the speed at which they run down an alley. Such skills are learned when a reward is made contingent on quantitatively constrained performance. Among human learners complex, precise skills (*e.g.*, tying shoelaces) are routine.

Discrimination learning. In discrimination learning the subject is reinforced to respond only to selected sensory characteristics of stimuli. Discriminations that can be established in this way may be quite subtle. Pigeons, for example, can learn to discriminate differences in colours that are indistinguishable to human beings without the use of special devices.

Concept formation. An organism is said to have learned a concept when it responds uniquely to all objects or events in a given logical class as distinct from other classes. Even geese can master such concepts as roundness and triangularity; after training, they can respond appropriately to round or triangular figures they have never seen before.

Principle learning. A subject may be shown sets of three figures (say, two round and one triangular; next, two square and one round, and so on). With proper rewards, the subject may learn to distinguish any "odd" member of any set from those that are similar. Animals as low in the evolutionary scale as the pigeon can master the principle of this so-called oddity problem.

The oddity problem

Problem solving. Examples of human problem solving are familiar: finding the roots of a quadratic equation, solving a mechanical puzzle, and navigating by the stars. Among other animals, chimpanzees have been observed to solve problems requiring toolmaking.

This list only samples from the remarkable array of animal activities categorized as learning. Beginning with habituation, they range from the simple adjustments of single-celled animals up to the highest intellectual accomplishments of mankind. It would be wonderful indeed if a single theory of learning were enough to account for all this diversity. So far, however, no theory of learning adequately covers more than a small fraction of these phenomena.

THE STATE OF LEARNING THEORIES

Yet, at the start of the 20th century, vast psychological systems, such as behaviourism and Gestalt psychology, indeed were offered as explanations of learning (and of much wider ranges of behaviour as well). And as late as the 1940s, comprehensive theories of learning were still believed to be reasonably near at hand. But during the next three decades it grew clear that such theories are tenable only for very limited sets of data. By the late 20th century learning theory seemed to consist of a set of hypotheses of limited applicability.

Important earlier theorists. Beginning in the 1930s a number of general theories were advanced in attempts to organize most or all of the psychology of learning. The most influential of the contributing theorists are noted below.

E.R. Guthrie (1886–1959) wrote that learning requires only that a response be made in a changing situation. Any response was held to be linked specifically to the situation in which it was learned. Guthrie argued that learning is complete in one trial, that the most recent response in a situation is the one that is learned, and that responses (rather than perceptions or psychological states) provide the raw materials for the learning process.

For E.C. Tolman (1886–1959) the essence of learning was the acquisition by the organism of a set of what he called Sign-Gestalt-Expectations. These referred to propositions said to be made by the learner that his own specific response to given signs (or stimuli) would result in such and such circumstances later on. Tolman seemed to be saying that what the learner acquires is a specific knowledge of "what leads to what." In brief, his theory was that the learner develops expectations based on experience and that learning depends entirely on successions of events. Although less vocal on the point than others, Tolman implied that learning was a gradual process.

The theory offered by Clark L. Hull (1884–1952), over the period between 1929 and his death, was the most detailed and complex of the great theories of learning. The basic concept for Hull was "habit strength," which was said to develop as a function of practice. Habits were depicted as stimulus-response connections based on reward. According to Hull, responses (rather than perceptions or expectancies) participate in habit formation, the process is gradual, and reward is an essential condition.

Comparison of these theories yields major questions for empirical investigation. Is learning continuous or discontinuous; is it a gradual or sudden (one-trial) process? Is learning a matter of establishing stimulus–response (S–R) connections or does it depend on the learner's understanding of perceptual relationships? Is reward necessary for learning?

Are theories of learning necessary? Such major investigators of learning as B.F. Skinner and J.A. McGeoch maintained in the 1930s and 1940s that preoccupation with theory was misguided. For them the approach simply was to discover the conditions that produce and control learned behaviour. Beyond this, their interests diverged. Skinner studied instrumental conditioning (operant conditioning, as he called it) among rats; McGeoch specialized in human rote memory. Although study of rote verbal learning had become heavily theoretical by the 1970s, Skinner and his associates stuck to their empirical guns, guiding a variety of programs for the practical control of behaviour. Teaching machines and computer-aided instruction, behaviour modification (*e.g.*, the use of tokens to reward desired behaviour among psychiatric patients), and planned utopian societies (Walden II) all found scientific origins in Skinner's rejection of theory in favour of direct efforts to produce results.

Intervening variables and hypothetical constructs. Learning is a concept and not a thing, and the activity called learning is inferred only through behavioral symptoms. The distinction implicit here between behaviour and inferred process is one of Tolman's major contributions and serves to reconcile influential views that might seem completely at odds. Classical behaviourism, as developed by John B. Watson (1878–1958), rejected every mentalistic account and sought to limit analysis to such physiological mechanisms as reflexes. Watson argued that these are objective in a way that so-called thoughts, hopes, expectancies, and images cannot be. The opposing view holds that experiential (introspective) activity (exactly what Watson sought to dismiss) does require discussion.

Watson's reflex theory

Tolman called himself a behaviourist and ostensibly was bound by Watson's insistence on objectivity. But he also was interested in thinking, expectancy, and consciousness. Tolman found his solution to this problem of incompatible theories after his association with the Vienna Circle of Logical Positivists, whose deterministic teachings he brought to the attention of U.S. psychologists about 1920. He maintained that learning is inexorably produced (determined) by such independent (directly manipulable) variables as the organism's previous training and physiological condition and by the response the environment requires. According to Tolman, the development of learning is revealed through the changing probability that given behaviour (the dependent variable) will result. He held that learning itself is not directly observable; it is an intervening variable, one that is *inferred* as a connecting process between antecedent (independent) variables and consequent (dependent) behaviour.

An attractive possibility is that intervening variables may have discoverable physiological bases. Psychologists Paul E. Meehl and Kenneth MacCorquodale proposed a distinction between the abstractions advocated by some and the physiological mechanisms sought by others. Meehl and MacCorquodale recommended using the term *intervening variable* for the abstraction and *hypothetical construct* for the physiological foundation. To illustrate: Hull treated habit strength as an intervening variable, defining it as an abstract mathematical function of the number of times a given response is rewarded. By contrast, Edward L. Thorndike (1874–1949) handled learning as a hypothetical construct, positing a physiological mechanism: improved conduction of nerve impulses.

Intervening variables and hypothetical constructs need not be incompatible; Thorndike's hypothetical neural process could empirically be found to be the mechanism through which Hull's abstraction operates.

Miniature theories. With growing realization of the complexity of learning, the grand theories of Guthrie, Hull, and Tolman generally have been abandoned except as historic landmarks. Hope for any impending, comprehensive theory was almost dead in the 1970s. More modest miniature theories remain, many likely to be of temporary value. An account of their major themes and issues, however, should have more enduring interest.

MAJOR THEMES AND ISSUES

Association. A dominant ancient theme in theories of learning has been that of association. Although the concept was accepted by Aristotle, it was brought into the developing psychology of learning by British empiricist philosophers (Locke, Berkeley, Hume, the Mills, and Hartley) during the 17th, 18th, and 19th centuries. Popular acceptability of the notion of association was related to progress in the physical sciences. The physical universe had been shown to consist of a limited number of chemical elements that can combine in innumerable ways. By analogy, a science of "mental chemistry" seemed appealing. The theorized elements in this new "science" were called ideas, said to be based on what were named sensations. The synthesizing principles by which these posited ideas combined in conscious experience were expressed as so-called laws of association. It was suggested that such conditions as temporal and spatial contiguity, repetition, similarity, and vividness favoured the formation of associations, and each was called a law of association. Thus, there were "laws" of repetition, of similarity, and so on.

Rote learning

At the end of the 19th century the notion of association was widely accepted among psychologists. German psychologist Wilhelm Wundt (1832–1920) took a position nearly identical with that of the British empiricist philosophers. Also in Germany, Hermann Ebbinghaus (1850–1909) began to study rote learning of lists of nonsense verbal items (*e.g., XOQ, ZUN, ZIB*). He maintained that the association of each word with every succeeding word was the primary mechanism in learning these lists. Pavlov in Russia offered temporary associative connections in the nervous system as a hypothetical basis for conditioned reflexes.

These European influences coalesced in North America. Wundt's notions were introduced there when a student of his from England, Edward Bradford Titchener (1867–1927), came to teach at Cornell University in Ithaca, New York. Ebbinghaus' method and theory became standard in Canadian and U.S. studies of verbal learning; Watson and other behaviourists applied Pavlov's conceptions to their learning experiments. Experimental psychology in the Western Hemisphere came to be dominated by what seemed to be a search for laws of association.

What is associated? Investigators asked whether associations are formed between observable stimuli and responses (S–R) or between subjective sensory impressions (S–S). One group that included Hull, Guthrie, and Thorndike took the relatively objective S–R position, while Tolman and others favoured the more introspective, perceptual S–S approach. For a time S–R theorists held popularity; behavioral responses are readily observable evidence of learning, and many included them in the associative process itself.

But the reduction of learning to mere external stimuli and overt responses raised discordant theoretical objections that the inner activities of the organism were being ignored. S–R theories failed to account for a host of learned phenomena. For example, people could be trained to say they heard sounds even when such auditory stimuli were absent. They said they dreamed about what they had learned, too; yet there need be no immediate external stimulus, nor does the dreamer always make the responses he dreams about.

Physiological psychologists and biologists found ways of delivering electrical stimulation directly to the brain; this eliminated the sensory stimuli and vocal or motor responses on which S–R theories hinge. Direct neural stimulation was found to be an adequate signal and the electrical response of the brain itself proved susceptible to conditioning. At this level of the nervous system, distinctions between stimulus and response mean less than at the periphery, and the S–S versus S–R controversy is no longer such a burning issue.

Direction of association. Classical conditioning dependably has been shown to proceed only forward in time. Bell must precede food if a conditioned reaction is to be established. If it had any effect, the reverse procedure (food before bell) would be called backward conditioning; but at most it only inhibits other reactions. There seems to be a relatively brief optimal interval in classical conditioning at which associations are most easily made. For quick reflexes such as the eyeblink, this interval is about one-half second; longer or shorter intervals are less effective. For slower reactions such as salivation the interval is longer, perhaps two seconds or so.

Forward and backward association; context association

In learning verbal associations the situation appears to be quite different. When one learns the Russian–English forward association *da*–"yes," he also learns the English–Russian backward association "yes"–*da*. Moreover, timing is much less critical than in classical conditioning. Verbal pairs are learned with almost equal ease whether presented simultaneously or separated by several seconds.

In what is called context association, the general environment may begin to elicit a response that is being conditioned to a specific stimulus. Thus, a dog may salivate simply on being brought into the experimental room—before any bell rings. Verbal associations also can be weakened by changes in the general situation.

Repetition. A major theoretical issue concerns whether associations grow in strength with exercise or whether they are fully established all at once. Evidence is that learning usually proceeds gradually; even when a problem is solved insightfully, practice with similar tasks tends to improve performance. Some (perhaps most) learning theorists have concluded that repetition gradually enhances some underlying process in learning.

Stimulus sampling, a statistical learning theory

The view that associations develop at full strength in a single trial leads to a typical question. How can the gradual nature of most learning be explained if all-or-nothing is the rule? One possible answer suggested by Guthrie has led to so-called stimulus-sampling theory. The theory assumes that associations indeed are made in just one trial. However, learning *seems* slow, it is said, because the environment (context) in which it occurs is complex and constantly changing. Given a changing environment, the sample of stimuli will differ from trial to trial. Thus, it is reasoned, it should take many trials before a response is associated with a relatively complete set of all possible stimuli.

In this light, the strength (or probability) of a response should increase with practice even if the elementary associative process occurs in a single trial.

These stimulus-sampling notions translate easily into mathematical form; they are an example of statistical learning theory, a more general development in the quantitative treatment of learning.

Reinforcement. Repetition alone does not ensure learning; eventually it produces fatigue and suppresses responses. An additional process called reinforcement has been invoked to account for learning, and heated disputes have centred on its theoretical mechanism.

Objectively reinforcement refers to the use of stimuli that have been found to facilitate learning. Under appropriate conditions, these include praise, food, water, opportunity to explore, sexual stimuli, money, electric shock, and direct brain stimulation.

More theoretically, the term reinforcement expresses various theoretical hunches about some specialized subjective quality all such stimuli might share. Food for a hungry animal is a well-established reinforcer, conceivably through its distinctive appearance and odour. It tends to elicit a set of responses: approaching, chewing, tasting, swallowing; these may produce additional perceptual activities that reduce the drive or desire for food (*e.g.*, by halting stomach contractions that are experienced as hunger pangs). But no single subjective quality imagined by theorists seems invariably effective in reinforcement studies. Perhaps some combination of introspective influences is critical, or it may be that perceptual processes apply differently from one learning situation to another.

Anti-associationistic positions. Not all psychologists have accepted the general validity of association theories; many have suggested that considerations other than association are crucial to learning.

Organization. Major critics of association theory included such Gestalt psychologists as Wolfgang Köhler (1887–1967), who held that learning often entails a perceptual restructuring of environmental relationships. Köhler cited his own studies of insightful learning by a chimpanzee. The animal learned to join two sticks (akin to a jointed fishing pole) as a tool to pull in a banana that was out of arm's reach and of either short stick alone. The ape was described as sitting quietly (as if in thought), and then suddenly fitting the sticks together to rake in the fruit. It was argued that the ability to perceive new ways of relating the sticks to the banana was essential in solving the problem.

Learning as perceptual organization

Similar organizational processes in perceiving can be demonstrated in serial verbal learning. Memorizing the list *thick, wall, it, tea, of, myrrh, seize, knots, trained* should demand some rehearsal. Yet, notice the phonetic resemblance to Shakespeare's famous line from *The Merchant of Venice*: "The quality of mercy is not strained. . . ." With that kind of perceptual organization, learning can become quick and easy.

A powerful argument also was made by psycholinguists who criticized what they took to be the associationistic account of language learning. Even assuming one-trial acquisition, it was held that such individually learned associations could not account for all combinations of words people use; there are simply too many. They suggested that learning a language requires some general organizing structure on which words are hung. Some proponents of this position hold that this structure does not depend on learning, being transmitted genetically from parent to child.

Inhibition. Gestalt interpretations often reject the associationistic hypothesis wholesale. Other theorists endorse the notion of association, but hold it to be less important than is a process of inhibition through which errors in learning are eliminated. Such theorists find support in evidence for the development of learning sets (what is called learning to learn).

For example, a monkey may learn a long series of discriminations; *e.g.*, red versus green, black versus white, round versus square, large versus small, triangle versus ellipse. After solving several hundred such problems, some monkeys learn to master each new one in a single trial, as if insightfully. The animal is said to have learned to learn such discriminations.

Learning to learn: error-factor theory

Evidence clearly shows that the monkey gradually abandons erroneous tendencies as learning proceeds. At first it might be prone to choose stimuli that are red, black, round, large, or triangular. Correct choices do not always correspond to the animal's initial biases, and their suppression (inhibition, extinction) eventually permits single-trial learning. Theoretically, organisms learn to learn by inhibiting erroneous behaviour; thus, Harry F. Harlow, a proponent of this view, called it an error-factor theory.

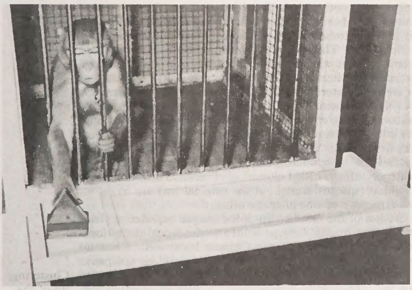
Motivation in learning. Motivation popularly is thought to be essential to learning. Yet many theorists suggest that motives make little or no direct contribution—that they simply tend to promote practice.

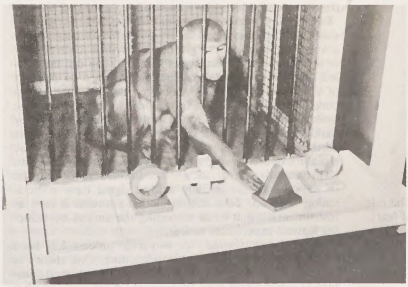
Motivation and performance. Learning was defined above as a change in a behavioral potentiality. Realization of such potential seems to be related to the learner's level of motivation. A pupil who has learned the names of all members of the British Commonwealth of Nations would be expected to recite them with particular energy under some sort of incentive (reward or punishment). The incentive is said to raise his level of motivation.

Incentives do seem to invigorate performance up to a point; however, when motivation seems particularly intense, some studies show performance to deteriorate. From such data some theorists conclude that the effect of drive intensity on performance follows a U-shaped course, first helping and later hindering.

Greatly increased motivation also may change performance qualitatively by introducing new inefficient modes of behaviour. A student may be so tautly driven to do well on an examination that his tension, fear of failure, and his visceral and muscular discomfort interfere with performance.

Motivation and learning. To show that motivation af-





Learning-to-learn experiment by Harry F. Harlow.
The monkey has been trained to expect food after lifting an object. In the first picture, the monkey lifts the triangle after having been shown a sample triangular form. In the second picture, taken several minutes later, the monkey selects the triangle from among other forms. The process is repeated with the circle and the cross.

fects performance of what has been learned is not the same as demonstrating its effect on the process of learning itself. This would require that individuals learn under various levels of motivation and be tested under the same incentive levels. (This is to control for the effects of motivation on performance alone.) And, indeed, the best-controlled experiments of this design indicate learning effects to be the same under different levels of motivation.

Varieties of learning. It is debated whether all forms of learning represent the same process. This question applies even to relatively primitive phenomena such as classical and instrumental conditioning.

Classical and instrumental conditioning

In instrumental conditioning reinforcement is contingent on the learner's response; a rat receives food only if it presses the lever. In classical conditioning there is no such contingency; a dog is fed whether or not it salivates. But this is a distinction in experimental procedure. Whether the underlying process of learning is the same for both is quite another question.

Classical conditioning usually has been reported for glandular, autonomically mediated, involuntary responses (*e.g.*, salivation, heart rate). By contrast, voluntary movements of skeletal muscles more typically have been found to be conditionable instrumentally. However, to theorize that classical conditioning is exclusively effective for one class of responses while instrumental conditioning is uniquely applicable to others seems to be a mistake.

Evidence that seems to demolish such theorizing comes from a series of experiments directed by Neal E. Miller at the Rockefeller University in New York City. Rats were immobilized with curare; this drug blocks the junction between muscle and nerve to paralyze the skeletal muscles. However, a curarized individual still can show autonomic, involuntary signs of emotional activity such as a rapidly beating heart.

Electrical stimulation of selected parts of the brain seems to be rewarding; animals behave as if they seek such stimulation and will learn to press a switch for it (voluntary muscle function). Using curarized animals, Miller and others made the rewarding stimulation contingent on such typically *involuntary* responses as changes in heart rate, blood pressure, contractions of the bowel, and salivation. Their research has shown such instrumental conditioning to be effective for all these responses. The evidence appears to destroy the once-popular hypothesis that involuntary autonomic reactions are subject only to classical conditioning. In this sense the two primitive forms of learning seem to be the same.

Stages of learning. Should the basic process prove to be the same for all varieties of learning, there would still be reason to believe that it operates differently from one stage of practice to another. For example, in coping with painful stimuli (*e.g.*, electric shocks) laboratory animals seem to learn in two successive, distinguishable phases. Apparently they first learn to fear the situation, *then* to avoid it.

For example, when an animal learns to avoid painful shock (by turning a paddle wheel or by running away), a warning signal can be given; *e.g.*, with a flash of light or a buzzer. The two stages of learning then can be studied separately. The animal first is subjected to pairings of signal and unavoidable shock to establish (by classical conditioning) signs of fear in response to the signal. In the second stage it is allowed to stop the frightening signal by making an appropriate response. Preconditioned members of the many animal species have learned to avoid the signal itself, even though shock never was presented again.

The role of fear

Theoretically, the classically conditioned signs of fright in response to the initially neutral signal have a motivating function. Termination of that stimulus is seen as instrumental—that is, as rewarding the animal by reducing learned experiences of fear.

Classical conditioning. A two-stage process has been suggested even for classical conditioning. One theory is that in the first stage the subject learns that a neutral stimulus (a ringing bell) is to be presented along with another stimulus (food) whether or not it exhibits a reaction (salivation). Conditioning of any reaction is held to constitute the second stage of learning. The skimpy supporting evidence points to the first stage as a prerequisite, suggesting that responses can only be conditioned after the sensory conditions are recognized.

Verbal learning. Theories that interpret verbal learning as a process that develops in stages also have been worked out. In one variety of rote learning the subject is to respond with a specific word whenever another word with which it has been paired is presented. In learning lists that include such paired-associates as *house–girl, table–happy,* and *parcel–chair,* the correct responses would be *girl* (for *house*), *happy* (for *table*), and *chair* (for *parcel*). By convention the first word in each pair is called the stimulus term and the second the response term. Paired-associate learning is theorized to require subprocesses: one to discriminate among stimulus terms, another to select the second terms as the set of responses, and a third to associate or link each response term with its stimulus term. Although these posited phases seem to overlap, there is evidence indicating that the first two (stimulus discrimination and response selection) precede the associative stage.

Remembering and forgetting. Learning, remembering, and forgetting often have been considered separate processes. Yet these distinctions seem to blur in the face of contemporary research and theory.

Transient and enduring memory. Evidence for stages of learning comes from observations of learners over relatively extended series of trials (or comparatively long periods). The empirical data suggest that several alterations in memory function occur even during a single trial. The process that commits information to memory also seems to have several stages.

Immediate, short-term, and long-term memory

Most theorists attribute at least three stages to memory function: immediate, short-term, and long-term. Immediate memory seems to last little more than a second or so. For example, subjects may be asked to remember where specific objects are located within a complex array they have just seen. Their performance shows that considerable information is retained only briefly, rapidly fading unless it is given special attention.

Short-term memory lasts about 15–30 seconds, as after looking up a telephone number. One makes the call, discovers he has forgotten the number (perhaps in the midst of dialing), and has to look it up again. Nevertheless, such short-term retention does make information available long enough to be rehearsed; if the learner repeats it to himself, the number can be transferred to some sort of longer term storage.

Thus, rehearsal seems to facilitate transfer of data from short-term to long-term memory. Once committed to long-term memory, the results of learning tend to endure but can be abruptly abolished when specific parts of the brain are injured or removed; they also are vulnerable to interference from other learning. Nevertheless, conditioned responses may undergo little or no forgetting over periods of months or years. And electrical stimulation of the surgically exposed brain while a person is awake can make him remember experiences long thought forgotten. Recall is reported to be similarly enhanced during hypnosis.

Retrieval. The amount of information one readily can retrieve from what is stored in memory is prodigious. In locating an item in memory, he apparently activates a system that stores a set of related data; then he searches for the item within that system. For example, a person is shown a long, randomly mixed list of words that belong to different categories (*e.g.*, names of animals, plants, professions, tools). When asked to remember as many words as he can, he spontaneously will tend to group them by category; this is called clustering of recall. Thus, names of animals (spread throughout the original list) are likely to be remembered one after the other.

Clustering; tip-of-the-tongue phenomenon

Studies of the familiar tip-of-the-tongue experience yield analogous results. College students who heard definitions (of this sort: a small, open Chinese boat) were asked to supply the right word (in this case it would be sampan). Those who said they might have it somewhere on the tip of the tongue were significantly accurate in guessing the first letter and the number of syllables. Their tendency also to recall words that sounded the same or that had similar meanings is reminiscent of clustering.

Considerable evidence of this kind supports the theory that the process of retrieval first locates stored data in some sort of associative network and then selects an item with specific characteristics.

Forgetting. Whether immediate and short-term data simply decay or are lost through interference is a matter of controversy. However, evidence is clearer that interference affects retention of information in long-term storage. Retention of the word *happy* (learned as a paired associate of *table*) seems to be subject to the interference of a strong tendency to associate *table* with *chair*. Thus, the paired associate *table–happy* becomes more readily forgotten when followed by *parcel–chair* as the very next item in a list; this seems to help *chair* reassert its old tendency to be associated with *table*. In general, it is found that associations tend to interfere with or to inhibit one another. Interference deriving from earlier (and later) associations is called proactive inhibition (and retroactive inhibition). These two forms of inhibition commonly are accepted as major processes in forgetting, proactive inhibition being assigned greater importance.

Contemporary trends in learning theory. In the early 1930s the distinction between learned and inherited behaviour seemed clearer than it does now. The view that any bit of behaviour either was learned or simply developed without learning seemed straightforward. Studies based on these expectations led investigators to conclude that rat-killing behaviour among cats is learned rather than instinctive, that human fears are all acquired, or that intelligence is completely the result of experience. Learning theorists were saying then that most behaviour is learned and that biological factors are of little or no importance.

Forty years later this position seemed grossly untenable. The once-implied sharp distinction between learned and inherited behaviour had become badly blurred. For example, it has been found that the young of many animal species automatically will learn to follow the first large, moving, noisy object presented (as if it were their mother). This special form of learning is called imprinting and seems to occur only during a critical early stage of life. Among mallard ducklings imprinting is most feasible about 15 hours after hatching. During this period a duckling will imprint as easily on an old man or on a rubber ball as it will on a mother duck. Is this instinctive or learned behaviour? Manifestly it is both. The instinctive tendency to be imprinted is part of the duckling's biological heritage; while the object on which it is imprinted is a matter of experience. What is significant for learning theory is that the contribution of biology cannot be ignored.

Imprinting, a special form of learning

Learning theorists once ruled a number of concepts out of court on the ground that they seemed objectively unclean. Image, cognition, awareness, and volition, all are concepts that were denied acceptance on this basis. They sounded mentalistic, subjective, introspective, and unverifiable. Yet, in the late 20th century these were being given more serious scientific consideration.

For example, the concept of image in learning has begun to show real viability. It has long been reported that the more meaningful a list of words is, the easier it will be to learn. Degree of meaningfulness for a word may be defined by the objectively observed probability that people quickly can give another word in response. Using such empirical scales of meaningfulness, a reliable and substantial relationship has been found between meaningfulness and ease of learning. However, meaningful words also may evoke vivid images that subjects can describe when asked. When they do evoke such imagery, they seem to be learned and remembered even more easily. Thus, learning theory seems to be enriched when introspective data are used.

A final fault in much learning theory stems from earlier tendencies to use the laws of physics as a model. Theorists once sought general laws of wide applicability that tended to obscure differences among individuals. For example, so complete was Hull's faith in universal "laws" of animal behaviour, that he based his hypothesis about the optimal interval for classical conditioning in humans, other mammals, and birds on the pattern of nerve conduction in the optic nerve of the horseshoe crab. There was little concern even for species differences. Within the same species, individual differences were viewed as a mere nuisance; it was believed that, by studying many subjects and by computing averages, basic laws of learning could be found. However, so-called laws were developed in this way that failed to represent even one individual whose behaviour contributed to the average. More than any other consideration, this has led learning theorists to take a belated look at the importance of individual differences and species differences in learning.

(G.A.K.)

Psychomotor learning

Human psychomotor skills are organized patterns of muscular activities guided by changing signals from the environment. Driving a car and eye–hand coordination tasks such as drilling a tooth, throwing a ball, typing, operating a lathe, and playing a trombone are behavioral examples. Also called sensorimotor and perceptual-motor skills, they are studied as special topics in the experimental psychology of human learning and performance. In research concerning psychomotor skills, particular attention is given to the learning of coordinated activity of the arms, hands, fingers, and feet; the role of verbal processes (see below *Concept formation)* is not emphasized.

THE RANGE OF SKILLS

The term "skill" denotes a movement that is reasonably complex and the execution of which requires at least a minimal amount of practice. Thus skill excludes reflex acts. One does not become skilled at sneezing or at blinking the eyes when an object approaches. At the same time, it has become increasingly apparent to scientists investigating the performance and acquisition of motor acts that the performance of complex skills is closely linked to sensations arising from the things the performer looks at, sensations from the muscles that are involved in the movement itself, as well as stimuli received by other sensory end organs. Thus the term "sensorimotor skill," denoting the close relationship between movement and sensation in the acquisition of complex acts, is found within the research literature with increasing frequency.

Simple components of bodily skills. Most of life's skills are continuous and complex and contain a multitude of integrated components; however, these complex skills may be analyzed by examination of their component parts.

The performance of a skill may be broken down into several time intervals. Initially the performer must be attentive and alert enough to be receptive to some kind of sensory information, which may in turn lead toward some kind of motor act. For this to occur the performer usually becomes aware of some kind of stimulus that is intense or distinctive enough to be perceived as different from other sensory information.

Reception of stimulus and selection and initiation of response

As this cue occurs the performer then must make a decision to act or not to act; and this is generally dependent upon his past experience within similar situations and with similar stimuli, as well as upon his feelings about his personal capabilities. If he decides to act, the next thing that occurs is the selection of an appropriate motor response from the entire "collection" of motor responses that he has acquired.

In the laboratory, a subject's reaction time is the time between the presentation of some kind of stimulus and the performer's initiation of response. The individual's speed of reaction in such situations is dependent upon a number of variables, including the intensity of the stimuli; for example, a person will initiate a movement more quickly to increasingly louder sounds until a limit is reached. If too loud a sound is presented it will delay the onset of the movement and result in a longer reaction time. Similarly, a longer reaction time will be recorded in such experiments if the subject is aware that he will have to initiate a complex movement or if the subject must choose from among a number of stimuli before initiating a movement (*e.g.*, if he must move only if one of a number of various coloured lights is turned on).

Factors that affect quality of motor response

The quality of the movement then initiated is dependent upon a number of other factors, the precision of the act required, the past experience of the performer in similar skills, the speed of the movement, the force of the motor act, as well as body part or parts to be moved.

The efficient performance of many types of extremely simple motor skills may be limited by inherent response capacities built into the human nervous system. Finger tapping at more than ten times per second for example is not usually possible. A person's ability to keep a body part relatively steady may be disturbed by natural, regular, and rather predictable oscillation rates of the limbs, fingers, and of the total body (evidenced in measures of body sway).

Individuals vary greatly in their ability to exercise force with various body parts. At the same time, there are limits beyond which it is not realistic to expect humans to go when evidencing strength in various tasks.

Careful experimentation reveals that because of the complexity of the human motor system it is unlikely that an individual ever repeats an apparently similar movement in precisely the same way. Thus the acquisition of skill in a given task involves the performance of a reasonably consistent response pattern, which varies, within limits, from trial to trial.

Basic abilities that contribute to motor skills

It is a common observation that there seem to be a number of basic motor abilities that may underlie the performance of a number of life's activities. This subject was investigated rather extensively during the 1950s and 1960s. The intercorrelation of performance scores elicited from thousands of young adult subjects who participated in these investigations revealed that there are in fact a number of basic abilities that contribute to the efficient performance of both fine and gross motor skills. Although a detailed examination of these abilities is not possible here, in general it was revealed that there are five components of what might be broadly referred to as "manual dexterity," including fine finger dexterity, arm-wrist speed and aiming ability. Similar research has explored the manner in which performance scores group themselves in skills in which larger muscles are involved. It was found, for example, that there are several kinds of strengths, including static strength (pressure measured in pounds exerted against an immovable object); this is independent of what was termed "dynamic strength" (moving the limbs with force). Muscular flexibility and balancing ability were similarly dissected into several components. Thus discussion of a single quality in human movement is inaccurate. Rather, one should refer to several specific types of ability.

Other classifications of motor skills

Motor skills may also be classified by reference to the more general characteristics of the tasks themselves rather than by measuring and intercorrelating the scores elicited by human subjects. It is common, for example, to find the dichotomy "fine" and "gross" motor skill, in which the latter label is applied to acts in which the larger muscles are commonly involved, while the former classification denotes actions of the hands and fingers. One researcher has proposed a three-way classification system, separating human movements into "body transport movements," in which the total body moves in space; limb movements; and manual (hand) actions. In general, however, most skills incorporate precise movements of both larger and smaller muscle groups working in harmony. The basketball player must use his larger skeletal muscles to run and jump but at the same time must employ a fine "touch," evidencing accurate finger control, when dribbling the ball. On the other hand, when a person sits at a desk and writes, the large postural muscles that contribute to the writer's stability are invariably active.

Complex, integrated skills. Most of life's skills are not simple ones. They are rhythmic at times and almost always are composed of several integrated parts. Such skills are often controlled by the organization of visual information available to the performer, particularly during the early stages of learning. At the same time, the individual's ability to analyze the mechanics of a motor task, his verbal ability, and other intellectual and perceptual attributes may influence his acquisition of a skill.

Although psychomotor skills are widely distributed—*e.g.*, in military, athletic, musical, and industrial settings—such complex situations typically do not lend themselves to rigorous experimental research. Most scientists have found it more analytically useful to study psychomotor learning under controlled laboratory conditions. Measures of proficiency obtained in the laboratory reflect increasing accuracy and decreasing variability in a learner's performance as training progresses. If there is sufficient genetic aptitude, a person's mastery of a skill depends on his motivation to improve, on his receiving continuous information or sensory feedback about the adequacy of his performance during training, and on such factors as the rewarding effects of corrections made during successive practice periods. Skills are susceptible to inhibitory influences. The full extent of gains in proficiency often is masked by temporary losses and emerges only later, without additional practice sessions.

Factors in performance

Psychomotor habits are mediated primarily by the sensory and motor cortex of the brain and by the neural fibres (commissures) that connect the two cerebral hemispheres. According to the majority of theoreticians, learning proceeds (habit strength develops) as a mathematical function of the amount or duration of rewarded (reinforced) practice. The effects of associative and motivational factors are believed to combine mathematically by multiplying one another, while inhibitory and oscillation (variability) factors are thought to have subtractive effects. Despite theoretical and empirical progress, much remains to be discovered about the learning and performance of psychomotor skills, especially about the interrelationships among training variables, feedback contingencies, and human-factor variables.

(C.E.N./B.J.C.)

LABORATORY RESEARCH IN PSYCHOMOTOR LEARNING

Devices and tasks. Hundreds of electrical and mechanical instruments have been developed for research in psychomotor learning, but those commonly used number less than two dozen. In operating a device called a complex coordinator, the learner is instructed to make prompt, synchronized adjustments of handstick and foot-bar controls to match different combinations of stimulus lights. Another device, a discrimination reaction timer, requires that one of several toggle switches be snapped rapidly in response to designated distinctive spatial patterns of coloured signal lamps. In performing on a manual lever, a blindfolded subject must learn how far to move the handle on the basis of numerical information provided by the experimenter. With a so-called mirror tracer, a six-pointed star pattern is followed with an electrical stylus as accurately and quickly as possible, the learner being guided visually only by a mirror image. The operator of an instrument called a multidimensional pursuitmeter is required to scan four dials and to keep their indicators steady by making corrections with four controls of the type found in an airplane cockpit. On a rotary pursuitmeter the trainee's task is to hold a flexible stylus in continuous electrical contact with a small, circular metal target set into a revolving turntable.

The pursuitmeter and the mathometer

Also employed in such research is a selective mathometer, on which the subject's problem is to discover, with cues provided by a signal lamp, which of 19 pushbuttons should be pressed in response to each of a series of distinctive images projected on a screen. While using a star discrimeter, a person receives information about his errors through earphones; his task is to learn to selectively position one lever among six radial slots in accordance with signals from differently coloured stimulus lights. A trainee on a two-hand coordinator has to manipulate two lathe crank handles synchronously to maintain contact with a target disk as it moves through an irregular course.

Measurements. The tasks required by the above devices produce a substantial range of psychomotor difficulty. The elements of skilled behaviour are expressed as numerical scores; *e.g.*, correct response and error percentages, amplitude and speed of movement, hand or foot pressures exerted, time on target, reaction time, rate of response, and indices of time-sharing activity. Most of the behaviour thus recorded lends itself readily to mathematical treatment. Laboratory devices for studying psychomotor learning characteristically exhibit high reliability (*i.e.*, intra-task consistency) and yield scores of useful validity (extra-task correlation) in predicting such behaviour as performance in factory work and the operation of motor vehicles and aircraft. In other words, it would appear that perceptual-motor devices reliably measure what they are designed to measure, and they also tap a significant proportion of the abilities required in real-life situations. When properly maintained and used under standardized conditions and when the resulting measurements are treated by statistical methods, the above devices are prime choices for many applied and basic research programs.

PHENOMENA OF PSYCHOMOTOR LEARNING

Acquisition. Speed and accuracy in the majority of psychomotor tasks studied are typically acquired very rapidly during the early stages of reinforced practice, the average rate of gain tending to drop off as the number of trials or training time increases (Figure 1). Curves based on such measures as reaction time or errors reflect the learner's improvement by a series of decreasing scores, giving an inverted picture of Figure 1. Tracking scores from the two sexes are seen in Figure 1. Other devices have yielded more complicated functions—*e.g.,* S-shaped curves for complex multiple-choice problems on the selective mathometer (Figure 2). Most acquisition curves obey a law of diminishing returns as high levels of skill are approached. Data such as those from tracking and multiple-choice tasks can be explained by rational mathematical equations derived from theoretical models (see formulas and captions in Figures 1 and 2). Between them, these two equations describe psychomotor acquisition curves from a wide variety of learning situations and of trainees with less than a 2 percent average error of prediction. Contrary to lay opinion, stepwise plateaus of proficiency are seldom seen, not even in learning Morse code. The "natural plateau" is a phantom.

Generalization and transfer. The occurrence of the phenomenon of generalization is seen in the tendency of laboratory subjects conditioned to respond to a particular stimulus (*e.g.,* a light) to respond as well to similar stimuli beyond the original conditions of training. As differences along a physical continuum (*e.g.,* brightness) between the stimuli used in training and those encountered on test trials increase, the effects of generalization decrease until there may be no transfer from one situation to another. Alternatively, the more the two situations have in common, the greater is the amount of predictable transfer. Generalization (or transfer) may be based on temporal patterns of stimuli (*e.g.,* rhythms), spatial cues (*e.g.,* triangularity), or other physical characteristics (see below *Transfer of training*).

Similarity and performance

The measured effects of prior training on the performance of a subsequent task define the transfer of psychomotor learning. Although it may be similar, the latter task usually differs measurably from that originally practiced. A common example is the ability required of many automobile drivers to change easily from, say, a three-speed transmission with a horizontal gear lever on the steering wheel to a four-speed mechanism with a vertical floor-mounted gearshift. In laboratory tasks, the amount and direction of transfer effects are accurately predicted. In practical skills, transfer is more likely to take place between tennis and badminton, for example, than between swimming and football, and between cornet and trumpet than between piano and tuba. Similarity is not the only correlate of transfer, however, and empirical studies must take account of such factors as the amount of practice and the sequence of events in previous training.

Transfer effects may be positive, negative, or zero; *i.e.,* learning one task may facilitate, hinder, or have no observable influence upon performance of the next task. Flight simulators are designed to maximize the amount of positive transfer, often by ensuring high levels of behavioral similarity. Negative transfer effects (*e.g.,* reaching for the floor to shift gears when the lever is on the steering wheel) appear occasionally but tend to be easily overcome. Since transfer necessarily involves retention, the best schedules minimize forgetting by the inclusion of short time intervals between training and transfer.

The degree and amount of transfer are contingent upon such factors as number of common elements or principles, stimulus and response similarity, amount of predifferentiation training, the variety of learning-to-learn experiences, part–whole relationships, differences in intertask complexity, use of mnemonic aids, and the extent of proactive or retroactive interference. Transfer equations usually assume that the basic indices of performance for experimental and control groups will increase with practice, that the possible measures range from negative 100 percent through zero to positive 100 percent, and that the groups have been equated in aptitude or initial ability to learn before the experimental treatments are begun. Retroactive interference designs typically employ a sequence of original learning, interpolated learning, and relearning.

Retention. Learning is to acquisition as memory is to retention. Psychomotor retention scores indicate the percentage or degree of originally learned skill that is remembered or recalled as a function of elapsed time. Alterations of motor memory are reflected by changes in means, variances, and correlations between test results. In contrast to verbal behaviour, which is notoriously susceptible to forgetting through interference within a matter of seconds, mean scores for tracking and coordination skills recorded over periods ranging from two days to two years diminish scarcely at all. Yet, when intervals of three minutes to six weeks are interpolated between discrete responses

Effects of rest intervals

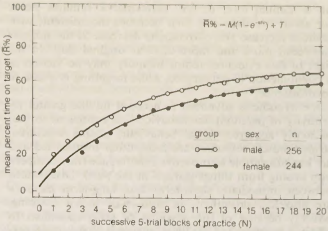


Figure 1: *Rotary pursuit acquisition curves.*
Average percent of response time on target ($\bar{R}\%$) while using a rotary pursuitmeter is plotted as a mathematical function of the number of successive 5-trial blocks of practice (N), with sex as the distinguishing factor, for 500 subjects. Both curves are predicted by the exponential equation shown in the figure, where *T*, *k*, and *M* represent origin, rate, and limit of theoretical response time, respectively.

From Clyde E. Noble, "Acquisition of Pursuit Tracking Skill Under Extended Training as a Joint Function of Sex and Initial Ability," *Journal of Experimental Psychology*, vol. 86, pp. 360–373 (1970); Copyright (1970) by the American Psychological Association, and reproduced by permission

From Clyde E. Noble, "The Length–Difficulty Relationship in Compound Trial-and-Error Learning," *Journal of Experimental Psychology*, vol. 54 (1957), pp. 246–252 (1957); Copyright (1957) by the American Psychological Association, and reproduced by permission

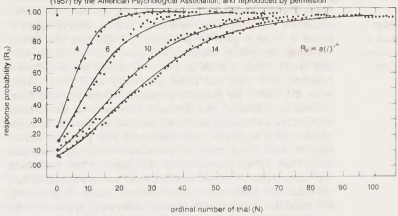


Figure 2: *Selective mathometer acquisition curves.*
Response probability (R_p), or relative frequency of correct multiple choices, in a selective mathometer learning experiment is plotted as a mathematical function of the number of successive practice trials (N), with task length (4, 6, 10, or 14 units) as the experimental condition, for 192 male subjects. All four curves are predicted by the double-exponential equation shown in the figure, where *i*, *r*, and *a* represent origin, rate, and limit of theoretical response probability, respectively.

on a manual lever device, performance remains stable for about two days and then becomes inconsistent; variabilities increase and correlations decrease as the subjects mis-recall more and more of their original skill. In the light of this evidence, motor memory may be viewed as a phenomenon of persistence, while forgetting is a case of inconsistence.

One hypothesis advanced to account for the greater retentivity of psychomotor behaviour, as compared to that of newly acquired verbal behaviour, is that nonverbal striped-muscle responses are more often overlearned and are less susceptible to proactive interference (*i.e.*, competition arising from things learned in the past). Distinctions between immediate, short-term, and long-term memory are also less prominent in studies of motor learning, possibly because of the devotion of skills specialists to efficient practice and feedback methods that ensure permanent storage of habits in the brain. This is not to say that motor skills are unforgettable; studies of short-term memory suggest that psychomotor forgetting can be swift indeed. Regardless of theoretical differences, however, psychologists generally agree that psychomotor behaviour is best remembered (and least forgotten) when overlearning is high, interference is low, reinforcing feedback is optimal, and interpolated activities are unrelated to the task being learned. Time is less important in the degradation of memory than are the events that fill the time (see also MEMORY).

Reminiscence. The phenomenon of reminiscence is a gain in performance without practice. Thus, when subjects performing trial after trial without rest (massed practice) are given a short break, perhaps midway through training, scores on the very next trial will show a significant improvement when compared with those of a massed group given no break. Reminiscence effects are most prominent in tasks demanding continuous attending and responding; they are least often observed with discrete-responding apparatus. The theoretical importance of this concept derives from its role in testing a hypothesis of reactive inhibition that asserts that a decremental process cumulates in the organism as a positive function of responding to stimulation and a negative function of resting time. That the phenomenon of reminiscence also manifests bilateral transfer of skill (*e.g.*, from the left to the right hand) suggests that the locus of the decrement is in the central nervous system rather than in the peripheral effector organs. Indeed, merely watching another subject practicing on a rotary pursuitmeter has an inhibitory effect on a person's performance; yet the cause is neither boredom nor fatigue.

Warm-up. Athletes and musicians often report that they get "cold" during a layoff (even for a rest period of a mere five minutes); when practice is resumed, the decrement in performance requires a warm-up before it is overcome. Similarly, on a rotary pursuitmeter, it is necessary to regain the optimal posture, grip the stylus correctly, begin the coordinated movements of eyes and hand, and recapture the proper whole-body rhythm. Warm-up produces a further gain in proficiency following the initial reminiscence effect. Mean scores continue to rise for several trials, reach a peak at the level found for distributed practice, and then fall more gradually until they merge with the curve for massed practice. When the duration of rest is extended, the amount of warm-up decrement first increases rapidly and then decreases; similar findings obtain for a succession of work and rest periods. Investigators who have tried to substitute warm-up activities other than actual pursuitmeter practice to offset or reduce the magnitude of the decrement have not been successful. At least for continuous psychomotor tasks of this sort, the need for proper, task-specific warm-up appears to be an intrinsic requirement of efficient performance. Wherever reminiscence goes, warm-up seems to follow; yet the converse does not always hold. The connection between warm-up and forgetting is uncertain.

Warm-up as a practice factor

Refractory period and anticipation. When required to make quick, discrete responses to two stimuli separated in time by one-half second or less, an operator's reaction time (latency) for executing the second response is typically longer than that of his first response. This difference in reaction time is called the psychological refractory period. At one time, it was thought possible that sensory feedback from the first response might stack up in the nerve centres to make the system refractory for a brief time, thereby delaying the processing of the second stimulus. Research findings that erroneous reactions could be corrected within one-tenth second would seem to negate the hypothesis. An alternative suggestion is that corrective movements are facilitated by feedback from the incorrect ones, and controlled observations appear to confirm that error-correcting responses have shorter latencies than those that are either correct or erroneous. Apparently, a false movement can be stopped on the basis of internal cues more promptly than on that of external stimuli.

Expectancy is a collateral factor with which researchers have had to reckon; *i.e.*, a subject may learn to accommodate himself to expect a delay between the first and second stimulus and thus be relatively unprepared should the second arrive earlier than usual. Further, people learn to be more expectant for particular kinds of stimuli than for others. When a person is uncertain about whether regularly occurring stimuli will be auditory or visual, or when their spatial direction is uncertain, performance is significantly degraded. This would suggest the possibility of divided attention; indeed, when pairs of stimuli are made perfectly predictable as to time and type, no impairment of response is observed.

Learning to divide attention

If a subject can acquire suitable expectancies via training and experience, then he can improve the skill of dividing his attention and, within physiological limits, simultaneously handle an increased range of stimuli without loss of proficiency. Results from extended practice on a task requiring successive choice and dual reaction indicate that, with learning, people can reduce the psychological refractory period. The ability to develop anticipatory responses to regularly occurring stimulus cues is well established. A military gunner scanning a distant fixed target for azimuth and elevation, for example, is engaging in a preview of receptor anticipation to maximize his score. An operatic tenor who rehearses covertly the opening notes of his cadenza while the orchestra finishes the introduction is employing perceptual anticipation to optimize his rendition. Anticipatory timing is learned, and reinforcing feedback is necessary.

FACTORS AFFECTING PSYCHOMOTOR SKILL

Amount of practice. It has been noted above (Figure 1) that the practice of sensorimotor tasks usually produces changes in scores that reflect diminishing returns. A major influence in learning generally, repetition is the most powerful experimental variable known in psychomotor-skills research. But practice alone does not make perfect; psychological feedback is also necessary. The consensus among theoreticians is that feedback must be relevant and reinforcing to effect permanent increments of habit strength. Once developed, habit never dies; it does not even fade away.

The effects of feedback and four other important performance variables (*i.e.*, task complexity, work distribution, motive-incentive conditions, and environmental factors) remain to be summarized.

Psychological feedback. Ranking prominently among experimental variables are so-called feedback contingencies (aftereffects, knowledge of results) that may be controlled by the experimenter so as to occur concurrently with or soon after a subject's response. A learner appears to improve by knowing the discrepancy between a response he has made and the response required of him; but, in experimental practice, the investigator manipulates behaviour by transforming functions of error. Since transformations are usually numerical or spatial, sensory returns from one's action may be informative, motivating, or reinforcing. Response-produced stimulation is intrinsic to most skeletal–muscular circuits; the neural consequences of bodily movement are fed back into the central nervous system to serve the organism's regulatory and adaptive functions. When this normal feedback is interrupted or delayed, psychomotor skill is often seriously degraded. Experimentally delayed auditory feedback of a subject's oral reading produces stuttering and other speech problems; delayed visual feedback in simulated automobile steering is a greater hazard under emergency conditions than is the driver's reaction time.

Generalizations about psychomotor learning

Laboratory investigations have supported the following generalizations about psychomotor learning: (1) without some kind of relevant feedback, there is no acquisition of skill; (2) progressive gains in proficiency occur in the presence of relevant feedback; (3) performance is disrupted when relevant feedback is withdrawn; (4) delayed feedback in continuous (but not discrete) tasks is typically decremental; (5) augmented or supplementary feedback usually results in increments; (6) the higher the relative frequency of reinforcing feedback, the greater is the facilitation of skill; and (7) the more specific the feedback (*e.g.*, in designating location, direction, amount), the better is the performance.

Experiments with a manual lever device, for example, suggest that when feedback is introduced and withdrawn at four stages of practice, the effect on error scores is profound. Knowledge of results given early and late has effects similar enough to reject any hypothesis that learning arises merely from repetition. These experiments indicate that practice makes perfect only if reinforced; the result of unreinforced practice is extinction of the correct response and a proliferation of errors. Studies employing a complex mirror-tracking apparatus have clarified the role of reinforcing feedback. Targeting performance was facilitated by presenting distinctive supplementary visual feedback cues previously associated with aversive (electrical shock) and nonaversive consequences. Moreover, the amount of facilitation grew curvilinearly with the number of cue conditioning trials. Work on human incentive learning thus demonstrates that the rate of gain in psychomotor proficiency can be regulated by stimuli that have been accompanied by positive or negative aftereffects. Persistence of the acquired reinforcing effects, considered with their cumulative quantitative properties, enhances the attractiveness of theoretical interpretations that emphasize continuity and reinforcement as contrasted with theories based on discontinuity and contiguity alone. Clark Hull's system (1943) is the classic model.

Human incentive learning

Task complexity. The complexity of discrete psychomotor tasks may be specified either as the number of response sequences a subject can make or as some measure of a subject's uncertainty about choices among stimuli. Still other factors that have been investigated as instances of complexity include variations in the number of possible responses at each choice point, different lengths of series, and regular versus unpredictable stimulus sequences.

Experimental procedures involving an increase of complexity produce more errors, require more trials to reach proficiency, and result in longer latencies per trial. Difficulty in psychomotor learning, therefore, generally increases with the complexity of the task to be mastered. An example of this phenomenon appears in Figure 2. Subjects exhibit continually altered probabilities of response during training sessions, and an average person with enough practice on a discrete sensorimotor task can learn to perceive, select, and react as fast to ten stimuli as he can to two. Apparently, it is not the number of choices among stimuli as much as it is the number of choices among responses that slows up a subject's processing activities and complicates his decision problems. Indeed, by limiting response alternatives (*e.g.*, circumscribing the physical range of a trainee's movements or providing supplementary auditory and visual indicators of error), a training device can facilitate the acquisition or transfer of skill.

Work distribution. Hazardous though generalizations about work and rest in psychomotor learning may be, a few guiding principles are notable: (1) massed practice is usually superior to distributed practice for simple discrete-trial tasks; (2) distributed practice is usually superior for complex continuous-action tasks; (3) short practice sessions are generally superior to long practice sessions; (4) long rest periods are generally superior to short rest periods, although forgetting must be counteracted; (5) for continuous-tracking tasks practiced under constant work sessions and variable rest periods, the final proficiency level grows curvilinearly as the intertrial interval is lengthened; (6) gains in proficiency under distributed practice, or with interpolated rest periods during massed practice, are usually in terms of performance rather than of learning; (7) losses in proficiency under massed practice, or with increased work load, usually pertain to inhibitory rather than motivational decrements; (8) under certain conditions (*e.g.*, "cramming" for examinations) it may be most efficient to mass practice as long as adequate rest can be obtained before criterion performance is demanded; (9) reminiscence increments and warm-up decrements are intimately related to schedules of work and rest; (10) decrement is not the same as fatigue.

Quite apart from the practical question of the optimal management of training programs (*e.g.*, in coaching oarsmen in racing shells), the aversive inhibitory consequences of sustained action that are recognized as subjective fatigue and behavioral decrement are clearly adaptive. By a reflex negative-feedback mechanism, inhibitory impulses may prevent an organism from working itself to exhaustion. With few exceptions, the presumption in favour of spaced practice can safely be taken out of the psychomotor-skills laboratory and applied in the gymnasium, lake, and playing field. Research on the skills involved in, for example, archery, badminton, basketball, golf, javelin throwing, juggling, marksmanship, rowing, and tennis supports the notion of distributing training by means of short workouts and frequent breaks.

Distributed practice applications

Motive-incentive conditions. Motivational processes are states of the organism that serve to activate reaction tendencies. Such states are classified as primary (innate) or secondary (acquired, learned) motivation. Though common physiological needs (*e.g.*, for food, water, avoidance of pain) may evoke psychological drives (*e.g.*, hunger, thirst, pain), the concepts of need and drive are not perfectly correlated. Some needs (*e.g.*, oxygen demand) seem to have no specific behavioral drive, and for some drives, clear-cut biological needs remain to be identified (*e.g.*, curiosity). Despite this apparent discrepancy, there is a theoretical consensus that psychological drive arouses the body to action, energizes its latent responses, and supports its behaviour over time. Most theorists believe that motivation (drive) and learning (habit) interact (in a multiplicative—drive times habit equals action—manner) in generating response. In other words, to produce action both are theoretically indispensable, but neither is sufficient alone. A person is not likely to perform a skill if he does not want to and cannot do so if he does not know what to do. The multiplicative theory implies that the same level of psychomotor proficiency may arise from quite different combinations of learning and motivation. Moreover, the organism's temporary drive state seems clearly to affect the adequacy of reinforcing feedback (*e.g.*, offers of monetary reward do little to arouse one who is already trying his level best). While these theoretical interpretations often apply well to laboratory animals, their application to human acquisition of skill is complicated because incentive learning in man can become very abstract.

Physiological explanations of human behaviour that depend on the concept of primary motives (derived from research with rats and dogs) run into difficulties in view of the fact that primary motivation and reward do not appear to be critical in most studies of human skill acquisition. Thus, instead of giving food pellets (as to a rat), an experimenter delivers praise to a human subject; rather than receiving feedback by electric shock, the human can be guided by a needle moving on a dial or a buzzer signalling an error. At any rate, despite efforts to distinguish such motivational factors as general drives from selective incentives, attempts to demonstrate significant motivational effects in human psychomotor learning have met with only modest success. Among exceptions to the above are a few studies with standard apparatus (*e.g.*, the complex coordinator) and with special devices that have indicated that such incentives as money, verbal threats, electric shock, exhortations, and social competition may be relevant. Significant effects frequently fail to appear in experiments, and findings are often contradictory, so it has been suggested that the intrinsic challenge of the

Contradictory findings about motivation

gadgetry, coupled with the subjects' already high pre-experimental motivation, leaves human volunteers unaffected by such weak laboratory manipulations of motive-incentive conditions as the foregoing (see also EMOTION, HUMAN; and MOTIVATION, HUMAN).

Environmental factors. Many practical skills must be executed outside the laboratory under unfavourable conditions of temperature, humidity, illumination, and motion. It is generally found that, below the limiting levels of extreme stress, such conditions affect psychomotor performance to a greater extent than they affect psychomotor learning. Representative findings have included the following: (1) isolation and sensory deprivation cause dramatic reductions in vigilance and monitoring skills within an hour; (2) environmental temperatures above or below 70 ± 5 °F (21 ± 3 °C) tend to lower scores on tracking apparatus but do not impair learning; (3) lack of oxygen slows reaction time, especially when the atmosphere corresponds to altitudes of 20,000 feet or higher; (4) accelerations of the body in a centrifuge or rotating platform disrupt postural coordination and produce systematic shifts in the perception of the vertical; (5) although such people as acrobats, dancers, pilots, and skaters can adapt well to high accelerations, even they lose equilibrium if deprived of the customary visual frame of reference; (6) rather mild centrifugal effects of slow, constant rotation may induce acute motion sickness and associated degradation of psychomotor proficiency in normal persons; (7) while some controlled work–rest schedules of crews during confinement in a small cabin upset daily sleep rhythms and lead to decrements in watchkeeping, memory, and procedural skills, a schedule of four-hours-on versus four-hours-off duty can be maintained for several months without significant impairment; and (8) faulty identifications of visual displays on an eye–hand matching task have been produced in volunteer subjects exposed to controlled infectious diseases (*e.g.*, respiratory tularemia, pappataci fever, viral encephalitis).

Other environmental stress variables found to exert negative influences are vibration, low illumination, high atmospheric pressure, noise, glare, toxic gases, ionization, and subgravity. Certain drugs have positive effects on psychomotor performance (*e.g.*, amphetamines, magnesium pemoline, methyl caffeine, pipradrol); some have deleterious effects (*e.g.*, alcohol, barbiturates, diphenhydramine hydrochloride, lysergic acid, meprobamate, phenothiazines, scopolamine, tetrahydrocannabinol, tripelennamine); and others are either neutral or have inconsistent effects (*e.g.*, caffeine, nicotine).

INDIVIDUAL AND GROUP DIFFERENCES

Statistical indices of psychomotor ability (*e.g.*, means, variances, and correlations) not only differ among individuals but may also serve to distinguish from each other groups of persons classified by such traits as age, sex, personality, and intelligence. Comparative psychological studies of monozygotic (identical, one-egg) and dizygotic (fraternal, two-egg) twins have indicated that high coefficients of heritability—which are measured as a ratio of genotypic to phenotypic variance—exist for perceptual, spatial, and motor abilities.

Types of changes with age

Age. The most pervasive differences in human performance on psychomotor apparatus are associated with chronological age, and scores obtained from nearly all the devices mentioned above are sensitive to age differences. Researchers generally report a rapid increase in psychomotor proficiency from about the age of five years to the end of the second decade, followed by a few years of relative stability and then by a slow, almost linear decrease as the ninth decade is approached. For simple hand or foot reactions, complex discrimination-reaction time, and coordinated automobile steering, the peak of skill is attained between ages 15 and 20 on the average. After this, performance declines, meaning that these driving skills at age 70 are generally on par with the skill levels of age 10. This is a two-stage process that starts with a developmental phase (through maturation) and is followed by the more gradual deterioration of aging. Common athletic skills—balancing, catching, gripping, jumping, reaching, running, and throwing—also improve through childhood; most athletes reach their prime before the end of the third decade. As the aging process continues, self-paced, leisurely sports like golf are favoured over opponent-paced, combative activities such as tennis.

Gender. Although the assessment of gender differences in perceptual and reactive abilities is complicated by a number of factors (including age and personality), girls and women tend to be more proficient than boys and men in such psychomotor skills as finger dexterity and inverted-alphabet printing. On the other hand, males generally do better than females at pursuit tracking, repetitive tapping, maze learning, and reaction-time tasks. On rotary pursuitmeter tests, women are not only less accurate but more variable than men of the same age (Figure 1). Although males appear to be superior to females in aptitude and capacity, these advantages disappear when subgroups are carefully matched for initial ability. In contrast, speed scores on discrimination-reaction tests reveal clearly diverging trends for college men and women trained intensively for several days (960 trials). Although both groups were equated for intelligence and had similar error scores, females showed cumulative impairment on the fourth day of practice, whereas males kept improving. Sizable differences in reaction latency and movement time are characteristic of the genders on other tasks.

Comparative strengths

Whereas girls tend to attain their maximum proficiency in speeded tasks earlier in life than boys do, males continue to gain proficiency over a longer period and maintain that proficiency well into middle age. After puberty, boys excel at athletic skills such as jumping, running, or throwing that demand stamina and strength. Differences in size, maximal oxygen consumption, percentage of body fat, and other physiological factors make it unlikely that even elite female athletes will surpass elite males, though they regularly outperform more than 99 percent of males. It should be noted that not all psychomotor differences associated with gender are intrinsic; unequal opportunities, distinctive social learning, role playing, and other cultural phenomena also influence the learning and execution of skills by males and females.

Environment and heredity. All humankind is of one species, yet psychomotor differences among human populations (just as those of morphology) can be identified. The Tarahumara far surpass others in long-distance running contests. Andeans and Tibetans are superbly adapted to working at high altitudes; Eskimos excel on psychomotor tasks performed under low-temperature stress. Given such examples, it is likely that the inherited factors underlying behavioral aptitudes and capacities may have evolved from different selective pressures in different ecological niches. As is true for age and gender, however, hereditary and environmental variables are complexly related.

Other factors. Many other characteristics contribute to psychomotor behaviour. The following, for instance, have been observed:

1. Speed scores in reaction-time tasks are positively correlated with body temperature in adults.
2. Psychotics show longer reaction times and poorer tracking scores than do people unmarked by psychosis.
3. Right-handed operators are favoured on the rotary pursuitmeter, while left-handed persons tend to do better on the complex coordinator.
4. Left-handed people are more variable in finger-dexterity and paper-cutting skills and also show more signs of ambidexterity.
5. Intelligence quotients (IQ) are weakly related to physical strength and endurance yet are strongly associated with performance in such activities as running the 35-yard dash, balancing on one foot, discrimination reaction, rotary pursuit, and selective mathometry.
6. Body build is associated with specific athletic skills—the best fencers, oarsmen, and basketball players, for example, tend to be tall and lean; top swimmers, divers, and pole-vaulters are likely to be broad-shouldered and slim-hipped; champion wrestlers, shot putters, and weight lifters are apt to be thick-trunked and short-limbed. While body type does not guarantee athletic prowess, it can contribute to success in certain sports.

Similar considerations apply to vocal and instrumental musical aptitudes wherein unique combinations of such anatomical structures as lips, teeth, larynx, tongue, eyes, ears, hands, and arms can facilitate the attainment of virtuoso skill.

In short, psychomotor abilities and learning underlie some of the most fundamental human activities, contributing to the full spectrum of work, play, creativity, love, and the very survival of the individual and the species.

(C.E.N./Ed.)

Concept formation

The process by which a person learns to sort specific experiences into general rules or classes is called concept formation. With regard to action, a person picks up a particular stone or drives a specific car. With regard to thought, however, a person appears to deal with classes: one knows that a stone will sink when thrown into water and that automobiles are usually powered by engines. In other words, these things are considered in a general sense beyond any particular stone or automobile. Awareness of such classes can help guide behaviour in new situations. Thus two people in a bakery may never have met before, but if one can be classified as customer and the other as clerk, they tend to behave appropriately. Similarly, most drivers are able to operate almost any automobile by knowing how to drive a specific automobile.

The term *concept formation* describes how a person learns to form classes, whereas the term *conceptual thinking* refers to an individual's subjective manipulation of those abstract classes. A concept is a rule that may be used to categorize a particular object within a certain class. The concept "citizen of the United States" refers to such a decision rule, meaning any person who was born in U.S. territory or who is a child of a U.S. citizen or who has been legally naturalized. The rule suggests questions to ask in checking the citizenship of any particular individual. As most concepts do, it rests on other concepts; "U.S. citizen" is defined in terms of the concepts "child" and "territory." Many scientific or mathematical concepts cannot be understood until the terms by which they are defined have been grasped. In this way concept formation builds on itself.

Discrimination learning

Conceptual classification may be contrasted with another type of classification behaviour called discrimination learning. In discrimination learning, objects are classified on the basis of directly perceived properties such as physical size or shape. The emphasis on concrete, physical features in discrimination learning can be contrasted with the more abstract nature of concept formation. When a stimulus is perceived to match several different past experiences, however, the response may be a compromise, because an object need not bear an all-or-none relation to a set of others in discrimination learning—for example, there is no absolute distinction between tall and short people.

While human beings are capable of abstract thought, many of the classifications people make seem to be concrete discriminations. For example, people may use the same term in a discriminative or conceptual way. A child might use the term *policeman* in discriminating a man who wears a distinctive uniform, while a lawyer may use the term to represent a male civil servant charged with enforcing criminal codes. In practice, people seem to think in ways that combine abstractness and concreteness. They also may blend class membership with assignment along a scale; *e.g.,* such concepts as leadership, an abstract quality that people are said to exhibit in varying degrees. The same would apply to vivacity, avarice, and other personality classifications.

People seem to develop more complex sets of classes than do other animals, but this need not mean that human modes of learning are unique. It may be that all animals have the same basic biochemical machinery for learning, but that human animals exhibit it in greater variety. Yet, it seems no more appropriate to account for human concept formation in terms of discrimination learning alone than it does to reduce the functions of a piston engine to chemical reactions.

EXPERIMENTAL STUDIES

Since careful observation of informal, everyday behaviour is difficult, most evidence about human concept formation comes from laboratory subjects. For example, each subject is asked to learn a rule for classifying geometric figures (see the Table).

Geometric Patterns of the Type Used in Studying Concept Formation

object number	size	colour	shape
1	big	green	triangle
2	big	green	circle
3	big	red	triangle
4	big	red	circle
5	small	green	triangle
6	small	green	circle
7	small	red	triangle
8	small	red	circle

The experimenter may concoct the rule that all green objects are called GEK. The subject is shown some of the figures, told which are named GEK, and asked to infer the rule or to apply it to other figures. This is roughly akin to teaching a young child to identify a class of barking animals with the name DOG. In both cases a general rule is derived from specific examples.

The problem of discovering that GEK = GREEN is almost trivial when four GEK and four NOT GEK figures are presented at once, but the problem becomes surprisingly difficult if the figures are presented one at a time and need to be remembered. Furthermore, when two concepts are to be learned together (*e.g.,* JIG = TRIANGLE and GEK = GREEN) memory for each concept tends to be mixed, and it becomes a formidable task to solve either problem. This suggests that short-term memory is important to concept learning and at the same time that short-term memory can often serve as a limiting factor in performance. The mastery of more complex concept learning often depends on allotting enough time for the information to be fixed in memory.

Concept identification

Most such experiments involve very simple rules. They properly concern concept identification (rather than formation) when the learner is asked to recognize rules he already knows. Adult subjects tend to focus on one stimulus attribute after another (*e.g.,* shape or colour) until the answer is found. (This represents problem solving with a minimum of thinking; they simply keep guessing until they are right.) People tend to avoid repeating errors but seem to make surprisingly little use of very recent, short-term experience.

Most people try out attributes in an orderly manner, first considering such striking features as size, shape, and colour, only later turning to the more abstract attributes (*e.g.,* number of similar figures, or equilateral versus isosceles triangles). This suggests that there is no sharp distinction between discrimination learning (relatively concrete) and concept formation (more abstract); instead, one progresses from the concrete to the abstract. If pairs of arrays are shown in which the same geometric figure is repeated (the rule being that all GEK arrays have exactly 10 figures), people will first react to the directly perceivable characteristics (*e.g.,* the extent to which the figures fill the space). They are likely to discriminate most grossly different NOT GEK patterns in this way quickly but will be troubled by arrays of 9 or 11. Eventually most adults should discover a solution by counting figures. Higher-order abstractions (*e.g.,* GEK arrays have an even number of figures), which ordinarily have a higher degree of difficulty, become easy to learn if the distinction is directly perceivable; for example, if even-numbered arrays are drawn symmetrically and all others are not.

Study can shift from concept identification to concept learning by requiring combinations of previously learned rules. A conjunctive concept (in which the rule is based on the joint presence of 2 or more features; *e.g.,* GEK patterns now are LARGE and GREEN) is fairly easy to learn when the common characteristics stand out. But learning a disjunctive rule (*e.g.,* GEK objects now are either LARGE or GREEN but not both) is quite difficult; there is no invariant, relatively concrete feature on which to rely.

Concept learning in adults may be understood as a two-step process: first the discovery of which attributes are relevant, then the discovery of how they are relevant. In the conjunctive illustration used here, the learner first is likely to notice that size and colour have something to do with the answer and then to determine what it is. This two-step interpretation presupposes that the subject has already learned rules for colour, size, shape, or similar dimensions.

Intra-dimensional shift

In an example of what is called intradimensional shift, initially the subject learns that GEK = GREEN; then, without warning, the experimenter changes the rule to GEK = RED. The same attribute or dimension (colour) is still relevant, but the way in which it is used has been changed. In extradimensional shift the relevant dimension is changed (*e.g.*, from GEK = GREEN to GEK = TRIANGLE) but the classification of some objects does not change (GREEN TRIANGLE is a GEK under both rules). The relative ease with which subjects handle such problems suggests something about how they learn. If they tend to learn simply by associating GEK with specific figures without considering the selected attribute, then they should find extradimensional-shift problems easier, since only some of their associations need be relearned. But if they have learned stepwise in terms of relevant attributes (*e.g.*, to say "What is the colour? . . . Ah, that colour means it is GEK"), intradimensional shift should be easier, since only the "how" phase of the two-step process need be relearned.

College students tend to find intradimensional-shift problems easier, indicating that they tend to use the two-step process. On the other hand, suppose a rat initially is rewarded when it runs into the right-hand side of a maze for food, then a change is made by rewarding entries to the left (intradimensional shift) or by rewarding entries to any brightly lighted alley regardless of location (extradimensional shift). The rat will perform best on the extradimensional-shift problem. Among children, performance depends substantially on age. Preschool children are likely to do best with extradimensional shifts (as rats do) but beyond kindergarten age, children tend to find the intradimensional shift easiest.

Concepts need not be limited to simple classifications. They can also be interpreted as models or rules that reflect crucial possibilities for change. To take a simple case, an adult is not apt to think that the volume of water changes when it is poured into a container of different shape. Young children may claim that it does. In the adult's concept, volume is not synonymous with the shape of a container but is based on a model of how fluids behave. Concepts offer a basis for deciding if certain changes will have significant effects.

AGE AND CONCEPTUAL BEHAVIOUR

Piaget's observations. Through clinical observations, Swiss psychologist Jean Piaget initiated considerable study of how young children learn concepts that help them cope with their physical surroundings. As models for defining feasible change, concepts are at least as important in such contexts as they are for classification. Piaget stressed that infants must first learn to distinguish themselves from the external environment. Next, they form understandings of the physical world (for example, identifying objects that fall) that allow further exploration of the world. Later in the preschool period, children grasp the concept of spatial localization—objects that are separated in space). Piaget characterized this period of learning as classifying objects only on the basis of perceptually attractive, concrete physical features (in agreement with laboratory studies of intradimensional and extradimensional shift).

He and others who used his methods reported that preschool children are apt to explain external change in terms of their own needs: a four-year-old is likely to say that a cloud moves "because the sun is in my eyes." Among children in early primary grades, other interpretations of cause and effect might be expressed by saying a moving cloud "wants to hide the sun." In later primary grades, volitional movement and passive movement usually become conceptually distinct. By adolescence, children develop an ability to deal analytically with objects apart from their immediate perceptual characteristics. This marks an understanding of the hierarchies of subclasses within more general classes—for example, a normal child of 11 applies the properties of all living things to the class called birds.

Given proper information, by age six many children display significant concept-forming abilities. They ordinarily have considerable linguistic competence, using (though often not being able to explain) such abstract qualifications as present and past tense. Rules of formal logic (such as "new math") can be taught in the elementary grades. Progressive use of abstract concepts seems to reflect both maturation and learning.

The role of instruction

The role of instruction in concept formation remains poorly understood, yet practically all cultural heritage is explicitly taught. How many people would develop the concept of number (let alone that of odd and even) if left to themselves? Human societies have existed for thousands of years without these concepts. Better knowledge of how to instruct and of the role of imitation in transmitting cultural concepts is needed. In addition, language itself guides how concepts will be formed; if a language has no words for a concept, it is unlikely that a speaker of that language will think of that concept.

Aging. It is generally thought that the ability to learn new abstractions tends to decrease in old age, while in extreme cases (such as senility, severe alcoholism, or brain injury), the deficit is dramatic. Much less is known, however, about changes in conceptual ability during the active period of adult life, in part because much of the evidence is conflicting.

As gifted children age, they tend to retain superior ability to grasp new abstractions. Among more typical people, however, little correlation is found between conceptual ability evaluated in the early teens and the same ability measured 10 or more years later.

In such abstract pursuits as pure mathematics or theoretical physics there is a tendency for creative scientists and writers to be most productive in their late 20s and early 30s, but there are many exceptions. As people get older they acquire a wealth of concepts that they can apply to a problem, so the net change in ability is hard to predict. Deterioration in learning new concepts is likely to be more rapid past 60, its severity varying markedly from person to person. Deterioration also may be caused by illness or injury. In general, fluid intelligence (the ability to manipulate abstract concepts) decreases with age while crystallized intelligence (the use of the accumulated concepts) increases with age. (E.B.H./Ed.)

LANGUAGE

When it is completely understood, the process of language acquisition should provide a window into the process of concept formation and human cognition in general. As a system of symbols abstracted from experience, language is a quintessential human characteristic. Even before the emergence of the field of cognitive science in the mid-20th century, theorists attempted to understand the issues at play. Benjamin Whorf, developing a theory expounded by his teacher Edward Sapir, famously posited what is known as the Sapir-Whorf hypothesis (or Whorfian hypothesis) in 1940, when he wrote:

The Whorfian hypothesis

> We dissect nature along lines laid down by our native languages. The categories and types that we isolate from the world of phenomena we do not find there because they stare every observer in the face; on the contrary, the world is presented in a kaleidoscopic flux of impressions which has to be organized by our minds and this means largely by the linguistic systems in our minds. We cut nature up, organize it into concepts, and ascribe significances as we do, largely because we are parties to an agreement to organize it in this way—an agreement that holds throughout our speech community and is codified in the patterns of our language. The agreement is, of course, an implicit and unstated one, *but its terms are absolutely obligatory;* we cannot talk at all except by subscribing to the organization and classification of data which the agreement decrees.

In short, Whorf held that the structure of a language conditions the way in which the speaker of a language thinks. Theorists still debate the meaning and feasibility of this hypothesis.

Another perspective was argued by the U.S. behavioral psychologist B.F. Skinner. He held that much linguistic phenomena can be explained by instrumental, or operant, conditioning—the same principles that describe how a rat learns to push a bar in response to a specific signal. Linguist Noam Chomsky countered that clearly children learn language (sounds, words, and syntax) without being explicitly taught, that some underlying structure for acquiring language must be species-specific.

Genetic theories

Chomsky and others hold that humans are genetically prepared to acquire some language at an early age, much as some birds show readiness to learn any song pattern to which they are exposed when they reach a certain age. In humans, this period seems to stretch from about age one to six. Should language prove to be a relatively independent biological function, the high linguistic competence of many young children with poor ability for abstract reasoning would appear less paradoxical.

Perhaps some rudimentary bases for language among other animals can be learned by methods appropriate for discrimination learning; even very young children are among the best discrimination learners in the animal kingdom. Once basic linguistic discriminations have been grasped, these can be used as tools with which the remainder of any language is learned. Linguists, psychologists, and theorists of many stripes continue to construct new tests and new combinations of theories in an effort to fully grasp the relationship between language acquisition and cognition. (Ed.)

CONCEPT FORMATION IN ANIMALS

As mentioned above, rats learn to enter lighted or unlighted alleys to get food, and goldfish can be taught to swim toward or away from an object. In such discrimination learning, the animal is said to associate a physical property of the stimulus with its response and with some contingency of reward or punishment. Thus, while a dog can be trained to come when called, it need not mean that he knows his name in the same sense that a human apparently does.

Most animals show classification behaviour that seems indicative of discrimination learning. A crow will respond to the danger call of a bird of another species—but only if that call resembles the crow's. Chimpanzees, however, have been observed using sticks as primitive tools; they behave as if they have a concept of things that extend reach. On considerable evidence of this sort, many are reluctant to say that animals are incapable of abstract thinking.

The oddity problem

Most studies aimed at evidence of concept formation among laboratory animals have involved primates, although there are reports of abstract behaviour among such animals as dogs, dolphins, pigs, and parrots. Monkeys have been taught to solve the oddity problem: presented with two objects of one kind and one of another, they can be trained to select the discrepant one. This behaviour persists even for sets of objects that have never been presented to them before. The animals behave as if they grasp the general concept of similarity, which is an abstraction rather than a simple discrimination. Animals have also been tested on the ability to learn languages. With great effort chimpanzees have been taught to speak and to use correctly a very few words. A much more successful attempt was made by Beatrice and Allan Gardner to teach a chimpanzee, Washoe, the sign language used by deaf people—gestures apparently being more appropriate to the anatomic structure of chimpanzees. (Human beings seem more prepared to learn spoken language.) The chimpanzee learned to use the signs for hat, dog, food, yes, me (self), sorry, funny, go, come, and many others. Moreover, Washoe's foster son learned 68 words simply through observation, while Washoe's companions learned to use sign language to communicate transactions and reassurance. They also taught others the concepts they had learned. (E.B.H./Ed.)

CONCEPT FORMATION BY MACHINE

Inculcating conceptual understanding in computers is far more difficult than originally envisioned by artificial intelligence (AI) pioneers. While computers are adept at handling deductive inferences (common in logic and mathematics), their mastery of inductive inference, or the formation of concepts based on experience, has proved more elusive. Conceptualizing, or "thinking," requires an immense web of information concerning "common sense" ideas of temporal sequence, spatial relationships, and cause-and-effect. The daunting problem of programming a computer with such general knowledge is known as the frame problem. In 1970 Marvin Minsky and Seymour Papert of the Massachusetts Institute of Technology (MIT) suggested working with simplified environments. An early example of this approach was Terry Winograd's computer program SHRDLU. Developed at MIT, SHRDLU interacted with a "virtual world" that contained blocks of various shapes, sizes, and colours, and a virtual robot arm to manipulate the blocks. Although many proponents of AI considered it a major breakthrough, Winograd soon announced that it was, in fact, a dead end—unable to scale up to the real world, SHRDLU had no idea what a small, red block was.

Much practical AI work has focused on expert systems, which employ specialized knowledge bases and inference engines that use production rules ("if-then" rules) elicited from experts in a particular field. In this restricted sense, machines have formed concepts. *See* ARTIFICIAL INTELLIGENCE. (Ed.)

Transfer of training

Will one's knowledge of English help him learn German? Are skillful table-tennis (Ping-Pong) players generally good court-tennis players? Can a child learn to multiply if he does not know how to add? These questions represent the problems of transfer of training: the influence the learning of one skill has on the learning or performance of another.

KINDS OF TRANSFER

Positive, negative, and zero transfer

Basically three kinds of transfer can occur: positive, negative, and zero. The following examples from hypothetical experiments, purposely uncomplicated by distracting detail, illustrate each. Suppose a group of students learn a task, B, in 10 practice sessions. Another group of equivalent students, who previously had learned another task, A, is found to reach the same level of performance on task B in only five practice sessions. Since the average number of practice sessions required to learn B was reduced from 10 to five, transfer of training from task A to task B is said to be positive ($10-5=+5$). Many successful training aids, such as those that simulate the cockpit of an airplane and that are applied to teach people how to use instruments for flying blind without leaving the ground, produce positive transfer; when students who have preliminary training in such trainers are compared to those who do not, those with preliminary training almost invariably require less practice in achieving the desired level of skill.

Sometimes the effect of transfer of training is to hamper effectiveness in subsequent activity. If after learning task A a group of people need 15 practice sessions to learn task B whereas only 10 sessions are required for those without any previous training in task A, then task A is said to lead to negative transfer of training on task B ($10-15=-5$). Having learned to drive on the right side of the road often is observed to produce negative transfer for the tourist from Japan or continental Europe or North America when he is traveling in Great Britain, where cars are to be driven on the left-hand side of the road.

The degree to which transfer of training occurs between two different tasks is often minimal and may be so small that it is called zero transfer. If learning task B with or without previous training in task A requires 10 practice sessions, then the amount of transfer from one task to the other is said to be zero ($10-10=0$). Learning to knit Argyle socks is apt to produce zero transfer of training in learning to sing an operatic aria in French.

Although in contemporary psychology transfer of training is a distinct topic of investigation with its own experimental designs and procedures for measurement, its implications pervade practically all of psychology, from conditioning to personality development. Ivan P. Pavlov discovered that, when a dog is conditioned to salivate in response to a sound wave of 1,000 cycles per second, it will also salivate if it is next exposed to a tone of 900 cycles per second, although typically the volume of saliva will be slightly reduced. In this case, transfer of training occurs between two similar auditory stimuli; in general, phenomena of this sort are called stimulus generalization. At the very root of modern theories of personality development is the assumption that what a person learns during his childhood will show a pervasive degree of transfer to his adult behaviour. In some cases stimulus generalization mediates this transfer. Some cases of excessive fears may have their origins in unpleasant experiences during early life.

EDUCATION AND TRANSFER

The experimental study of transfer of training has historical roots in problems of educational practice. Educators in Western countries at the end of the 19th century widely endorsed the doctrine of formal discipline, contending that psychological abilities, called "mental faculties" by such philosophers as Thomas Aquinas (1225–74), could be strengthened, like muscles, through exercise. By learning geometry, one was expected to improve his ability to reason; studying Latin was held to "strengthen" the so-called faculty of memory, and so on. Although what contemporary educators have demoted from the doctrine to the theory of formal discipline once seemed reasonable to many, experimental tests have refuted it. When the reasoning abilities of groups of mathematics students in secondary schools were compared with those of other equally talented students who had not had the same mathematical training, no differences in general logical effectiveness were observed between the groups.

An alternative theory of identical elements was proposed in which it was postulated that transfer between activities would take place only if they shared common elements or features. Thus it was predicted that one's training in addition would transfer to his ability to learn how to multiply. It was reasoned that both tasks share identical features, multiplication basically requiring a series of successive additions, and that both tasks demand the individual's concentration.

Theories of transfer

But the identical-elements formulation soon came under attack when experimental results suggested that one's understanding of general principles, rather than the presence of identical task elements, has substantial effects on transfer of training. In one notable experiment, two groups of boys practiced throwing darts at a target placed under about a foot of water. Only one group, however, was instructed about the principle that water bends (refracts) light. According to this principle, the apparent position of the target should vary with the depth of the water. When the target depth was reduced to four inches, the group that had been taught the general principle of refraction adjusted rapidly to the change and exhibited substantial positive transfer; the other boys showed comparative difficulty in learning to hit the target at the shallower level.

These formulations (formal discipline, identical elements, and general principles), when considered carefully, might be recognized as points of view rather than as rigorously specified theories that could lead to unequivocal predictions of the results of new experiments in transfer of training. For example, failure to demonstrate positive transfer between mathematical training and general reasoning ability could be attributed to ineffective teaching of mathematics; in such a case, the results need not be interpreted as refuting the theory of formal discipline. If the then-traditional manner of teaching mathematics could be changed to emphasize logical thinking (rather than routinized application of formulas), it was argued that perhaps mathematical training could improve reasoning ability in general. Some theorists also suggested that the positive transfer observed to result when boys learned the principle of refraction was consistent with the hypothesis of identical elements; these theorists observed that a general principle may be considered an element common to many tasks. According to this line of reasoning, the group of boys who exhibited positive transfer with the shift to a new target depth shared the principle of refraction as an element in common with the previous task, along with those of aiming and throwing. By contrast, the youngsters who performed without the benefit of knowing about refraction were held to have gained positive transfer from throwing but to have suffered negative transfer as a result of aiming incorrectly.

EXPERIMENTAL ANALYSIS OF TRANSFER OF TRAINING

The indeterminate character of the broad theoretical formulations offered to account for transfer of training and the often unsuccessful ways in which they were applied to the practical problems of classroom teaching led some psychologists to retreat to the laboratory in the hope of identifying more clear-cut, fundamental processes in transfer of training. As a result, a number of different transfer-of-training phenomena were discovered, several of which may be reviewed as follows.

Stimulus and response similarity. The method of paired-associate learning, in which a person is asked to learn to associate one syllable or word with another (*e.g., complete–hot, safe–green, wild–soft*), encouraged the investigation of the influence of stimulus and response similarity on transfer of learning. Typically these pairs of verbal items are presented to the laboratory subject so that the first, or stimulus, member (*e.g., complete*) is exposed alone, followed after a short interval by the second, or response, member (*e.g., hot*). The subject's task is to respond to the stimulus term before the response term appears, as when an English-speaking student in learning French is supposed to respond to *le livre* with *the book.*

When two successive lists of paired associates are learned in which the stimulus elements are the same but the response terms are changed (*e.g., complete–hot* in the first list and *complete–new* in the second), negative transfer typically results. Apparently, in learning the second list the subject tends to respond to the stimulus term (*e.g., complete*) with the previously learned correct response term (*e.g., hot*), the result being interference with new learning to produce negative transfer. If he were learning the second list without having learned the first, the subject would not be so handicapped.

Another question concerns the sort of transfer that results when response terms are different and stimulus elements are similar but not identical; for example, *entire* is similar to *complete.* After one has learned *complete–hot,* the experimental evidence is that his ability to learn *entire–new* becomes definitely more difficult. Both *entire* and *complete* seem to have a tendency to evoke the response *hot* and to be incompatible with subsequently learning the association of *entire* with *new.* The principle that appears to operate in such situations is that the greater the similarity in stimulus elements, the greater the degree of negative transfer.

The influence of response (rather than stimulus) similarity on transfer of training is more complex; in paired-associate learning, the subject needs to learn the response term of each pair (response learning) and then to remember that it is linked with its appropriate stimulus partner (associative learning). When response terms are relatively difficult to learn (as in the case of unfamiliar or foreign words), the subject tends to profit considerably from learning the first list. But when response terms already have been learned (or are easy to learn), little if any positive transfer is likely to occur. The degree of transfer between lists that contain similar response terms depends both on how similar they are and on their level of difficulty; increasing the similarity between response terms is most likely to increase positive transfer when the response terms are relatively difficult to learn.

Obstacles to a unified theory of similarity in transfer

Although attempts have been made to formulate an all-embracing theory that would account for the effects of similarity among paired associates on transfer of training, a major obstacle that has prevented fully satisfying results is that the degree of positive or negative transfer is typically a product of many interacting influences beyond those of stimulus and response similarity. For example, the amount of training that the subject receives also has significant effects on transfer. When initial training is given on a simulated task (*e.g.,* learning to operate a set of dummy controls in preparation for a second task of acquiring a complicated skill, such as flying an airplane), negative transfer effects frequently appear during the initial stages of learning the second task and then give way with further training to generally positive transfer effects.

Another stumbling block in developing theoretical explanations has to do with the meaning of the central concept of similarity. In such experiments as those in which the salivary reflex is conditioned to different auditory stimuli, similarity is measured in terms of physical stimulus properties (*e.g.*, pitch or loudness); in other studies, as in paired-associate learning, similarity typically is expressed in terms of verbal meaning. In neither case has a universally adopted method yet been devised to measure similarity in a reliable and precise way; perhaps none can be, simply since there are so many different aspects of physical and linguistic or semantic similarity. Despite these difficulties, efforts to analyze transfer experimentally in terms of the properties of stimulus and response events have been productive in identifying conditions that can be varied to alter the direction and the degree of transfer of training.

Retroactive and proactive inhibition. Closely related to stimulus and response similarity are phenomena called retroactive inhibition and proactive inhibition; these demonstrate how forgetting seems to result from interfering activities.

In a study of retroactive inhibition, both the experimental and control groups of people learn task A (for example, a list of adjectives) and are tested for their ability to recall A after a specified time interval. The groups differ in what they are asked to do during the interval; the experimental group learns a similar task B (say, another list of words), while the control group is assigned some unrelated activity (for example, naming a series of coloured chips) designed to prevent them from rehearsing task A. The results of numerous studies of retroactive inhibition show that the experimental subjects typically are poorest in recalling information from task A. The interpolated activity, particularly a comparable one such as memorizing a second list of adjectives, apparently interferes with one's ability to recall words from the first list. Habit competition, or what is sometimes called interference, between the items of the original and the interpolated word lists at the time of recall is considered to be one of the major sources of the negative transfer exhibited in retroactive inhibition.

Interference theory

Experimental designs for demonstrating proactive inhibition differ from those used for showing retroactive inhibition in that the experimental group learns task B before, instead of after, task A. Whereas B was a task that was interpolated between the learning and the recall of task A in the retroactive inhibition study, B is a task that precedes the learning of task A in the proactive inhibition study. To evaluate the effects on the experimental subjects of their having learned B prior to A, the control people are instructed to relax during the time the experimental group is learning B. Typically an experimental subject's ability to recall from task A is inferior to that of a control person, the degree of inferiority depending in part on how similar the two tasks are; the greater the similarity, the poorer the recall tends to be. Although proactive inhibition, so called to indicate that it acts forward from the first-learned task to the second, produces appreciably less forgetting than does retroactive inhibition, they both support the theory that interference can produce forgetting (see MEMORY).

Stimulus predifferentiation. Educational films can be considered as everyday examples of stimulus predifferentiation, in which the individual gets preliminary information to be used in subsequent learning. The student who sees a film describing the various parts of a microscope is likely to be better prepared to learn the requisite skills when confronted with the instrument itself. In laboratory studies of stimulus predifferentiation, the subject is given experience with a particular stimulus situation ahead of time; later he is asked to learn new responses in the same situation. In one study, subjects practiced labeling four different lights and then later were asked to learn to press selectively one of four switches, each connected to one light. The rate at which they learned the appropriate pressing reactions was related to how well they had learned to label the lights.

The results of a large number of experiments covering a variety of stimulus predifferentiation techniques suggest that, when a learner has an opportunity to become generally acquainted with an environment, he retains some information about its different components that prepares him for learning to make new responses to them. Various explanations have been offered to account for this facilitation; some investigators suggest that the process of labeling enhances the distinctiveness of environmental stimuli for the labeler; others hold that perceptual acquaintance can more sharply differentiate an environment into its component parts for the perceiver or that it may encourage appropriate responses of observing or attending. Nevertheless, no single process has been identified as fundamental in stimulus predifferentiation. Perhaps a number of these processes operate in different combinations from one stimulus-predifferentiation transfer experiment to another, each process representing a different method by which a learner can become familiar with the details of his environment.

Transposition. Another phenomenon that has received considerable attention in theories of transfer of training is called transposition. An initial report of transposition came from a study in which chickens were trained by rewards to respond to the darker of two gray squares. After this discrimination task was learned, the chickens were shown the originally rewarded gray square along with one that was still darker. They seemed to prefer the darkest gray to the square that had been previously rewarded. This finding was interpreted to support the hypothesis that the birds had initially learned to respond to a relationship (what a human being would call the concept "darker") and that this response to a relationship had been transposed or transferred to the new discrimination. This relational interpretation later was challenged by theorists who offered a formulation to show, on the basis of principles of stimulus generalization, how a response to a relational stimulus could be explained by assuming that organisms do indeed respond to the absolute properties of the stimuli. Both explanations were found to be too simple for the variety of findings obtained with transposition studies. As a result, the interest of many investigators shifted away from demonstrating the relative merits of absolute versus relational interpretations to identifying conditions that seem to influence transposition behaviour. Within this context, newer, more sophisticated formulations have been proposed that consider both the absolute and relational characteristics of the stimuli in transposition studies.

Transfer among animals

Learning to learn. When people are asked to learn successive lists of words, their performance tends to improve as they progress from one task to another so that much less time is commonly required to learn, say, the tenth list than was needed for mastering the first list. This improvement suggests that information beyond the specific content of lists of words is also learned. It would seem as if the subjects are learning how to learn; that is, they seem to be acquiring learning sets, or expectancies, that transfer from list to list to produce continually improving performance.

Some of the most intensive work on learning sets has been carried out with monkeys that were learning how to solve several hundred discrimination problems in succession. In each problem, the monkey learned which one of two objects (for example, a bottle cap and a cookie cutter) consistently contained a piece of food. Although the solution of each successive problem required the animals to discriminate between two previously unfamiliar objects, performance tended to improve on successive tasks; the monkeys made increasing numbers of correct choices on the second trial of each problem as the process continued. Manifestly there was no cue to indicate the correct choice on the first trial of any specific problem. If the animal responded correctly on the first trial, then on the second trial it would only have to choose the same object to be correct thereafter; if the monkey made an error on the first trial, then the other object would inexorably be the one that should be chosen next. During their efforts to solve the first few problems the monkeys were correct approximately half the time on the second attempt to solve each problem. This success increased to an average of 80 percent correct after each animal had solved 100 problems, to 88 percent after 200 correct solutions, and eventually to 95 percent after 300. Thus, after a long series of separate tasks, all of the same type, the monkey's first response to the next problem usually provided sufficient information for the animal to make the correct choice.

Since each of the successive discrimination problems was different, what actually was being transferred from problem to problem? In these discrimination problems, the monkeys seemed to have several items of information to learn in addition to which one of the two objects contained the rewarding bit of food. The animals apparently had to learn to pay attention to that part of their environment where the objects were placed. To make the correct choice, it would seem that a monkey would have to learn to abandon any preference it might exhibit for objects on either the left or the right; indeed, the animals usually did show such preferences. (The correct object was shifted from side to side in a random sequence to control for these preferences.) Ostensibly, the monkeys also had to learn that one object consistently contained food while the other was always empty. Although these learning sets by themselves would not serve to identify the correct object in each new discrimination problem, it seems likely that they could help the animal locate the reward very rapidly by eliminating initially unprofitable responses.

Reversal learning. In reversal learning, the individual first learns to make a discrimination, such as choosing a black object in a black–white discrimination problem, and then is supposed to learn to reverse his choice—*i.e.*, to choose the white object. Such reversals tend to be difficult for most learners since there are negative transfer effects; *e.g.*, the individual tends to persist in responding to the black object that was originally correct. Eventually, however, one's tendency to make the originally learned selection typically becomes weaker, and he makes the competing response (*e.g.*, to white) more frequently until a point is reached where it is almost consistently evoked. Reversal learning can be accomplished very rapidly when a laboratory animal, such as a monkey, is presented with a series of reversal-learning problems in which the same sequence of shifts is repeated (as when black is initially correct, then white, then black, then white, and so on). After extended reversal training, some animals are able to make the next reversal in the sequence in one trial. They behave as if they have mastered the abstract concept of alternation or of regular sequence.

Evolution and transfer

The speed with which representatives of a given species of animal, including human beings, can be taught to make a reversal of this kind seems to be related to the place biologists assign them in a hierarchy of evolutionary development. On first being exposed to a reversal-learning problem, normally competent adult humans who can use language are likely to achieve a solution with great rapidity. Monkeys can learn to perform equally well after a relatively longer series of reversal-learning tasks; but isopods such as pill bugs or sow bugs, small relatives of crabs and shrimp, have such primitive brains that they seem to be unable to improve their performance at all during a series of reversal-learning tasks.

DEVELOPMENTAL PROCESSES AND TRANSFER

The manner in which a problem is learned seems to have an effect on what is transferred. This conclusion is supported by experiments in which comparisons are made of the relative ease with which children of different ages execute reversal and so-called extradimensional shifts (see above *Concept formation*). In performing both kinds of shift, experimental subjects learn two successive discriminations between two pairs of objects that vary simultaneously in two aspects or dimensions—*e.g.*, white triangle versus black square, and black triangle versus white square. In training subjects initially, discrimination of only one dimension (for example, black–white) is made relevant, with the child's selection of one of the cues (for example, white) being rewarded, while the other (black) is incorrect. After they have learned this, the children are shifted to the second discrimination. In the case of a reversal shift, the same stimulus dimension (black–white) remains relevant, but the child is now to learn to reverse his initial choice; black choices are now rewarded, and white selections become incorrect. For an extradimensional shift, the initially irrelevant dimension (square–triangle) is given relevance by rewarding selection of one of its alternatives and by failing to reward choices for the other.

Age and transfer

The relative ease with which human beings learn to make extradimensional and reversal shifts is related to how old they are. Reversal shifts are relatively difficult for young children to learn and are relatively easy for adults to master. As people gain maturity, the relative ease with which they execute a reversal shift tends to increase in comparison with their ability to achieve an extradimensional shift.

Explanations for these developmental changes seem to be found in the manner in which the individual solves a discrimination problem. Very young children and laboratory animals tend to learn simple habits when faced with a discrimination problem for the first time; for example, they are most likely to learn simply to approach black objects and to avoid white. Reversal shift is often extremely difficult for them, and negative transfer effects are substantial. Subjects who primarily learn simple habits are faced with the task of eliminating one habit (*e.g.*, to choose black) that has been rewarded and then of developing another habit (*e.g.*, to choose white) that previously has not been rewarded.

Human adults, on the other hand, generally find a reversal shift relatively easy; they do not behave as if they simply associate their choices to the relevant stimuli (*e.g.*, white and black) but instead appear symbolically (or conceptually) to react to both of them in terms of their common characteristic (brightness). A similar kind of symbolic or logical response is appropriate in solving reversal-shift problems; since the relevant dimension remains the same, this kind of shift tends to be easier to make than is one involving extradimensional shift, which requires the individual to switch to a new symbolic response (*e.g.*, from brightness to size). In short, when they respond concretely, learners favour their potentials for achieving extradimensional transfer; those who tend to respond symbolically enhance the probability for reversal transfer.

Whatever the validity to be found in theoretical explanations of this sort, review of how transfer phenomena may be influenced suggests that no single principle or simple theory thus far put forward accounts for all of the observed data. Instead, the evidence is that several interacting processes underlie transfer of training and that their relative influence depends both on the nature of the tasks between which transfer takes place as well as on the characteristics of the learning organism. If one seeks to control the degree of transfer, as one does in educational settings, it seems useful to analyze transfer behaviour in terms of a number of component processes—*e.g.*, stimulus and response similarity, stimulus predifferentiation and response learning, and the symbolic abilities of the learner.

THE PHYSIOLOGY OF TRANSFER OF TRAINING

Cross education

Although available evidence for a physiological basis of transfer of training is limited, some impressive data already are recorded. Some central (brain and spinal-cord) mechanisms seem to control transfer of training. A long-established transfer phenomenon is cross education, in which there is positive transfer of a skill learned with one part of the body to another, untrained part. For example, a person who learns to throw a dart with his preferred hand exhibits positive transfer to his non-preferred hand. Since different muscles are involved in the equivalent action of opposite limbs, positive transfer resulting from cross education cannot be attributed simply to common muscular movements; instead it would seem that cross education depends on central processes that control the actions of both limbs.

Among highly evolved animals, transfer of training between limbs from opposite sides of the body evidently is mediated through a massive system of neural fibres, known as the corpus callosum, that connects the two hemispheres of the brain. One of the many ways in which the validity of this principle may be demonstrated is first to train blindfolded cats to discriminate with one paw between two different pedals (by feeling raised horizontal lines on one pedal and by detecting raised vertical lines on the other). Since each eye sends some of its nerve impulses to both hemispheres of the cat's brain while each paw only directs impulses to the hemisphere of the brain on the same side of the animal's body, this procedure feeds the sensory information to just one hemisphere. After learning to make the discrimination with one paw (*e.g.*, reward being given only for the pedal with the horizontal pattern), a cat that is confronted with making the same discrimination with the other front paw, which has its connections with the ostensibly "untrained" brain hemisphere, will nevertheless exhibit positive transfer. Indeed, even when the corpus callosum is surgically severed immediately after learning (to "disconnect" the two hemispheres), positive transfer will take place from one front paw to the other; manifestly, transfer of training takes place between connected hemispheres while the animal is learning. If the cat's corpus callosum is severed before it initially learns to discriminate the two pedals, however, no transfer occurs between the animal's limbs; the untrained paw fails to exhibit any benefit from what has been learned with the other paw. In other words, by severing the cat's corpus callosum, the surgeon splits the brain into two independently functioning units. The same kinds of behaviour are observable among other split-brain animals, including chimpanzees and people.

The physiological foundations of transfer of training are not limited merely to the anatomical considerations of the central nervous system. To better understand how physiological processes mediate transfer of training means also to be able to specify more fully the anatomic, electrical, and chemical basis of learning in general, a goal that remains incompletely achieved. Many physiologists and psychologists hold that the search for the neurophysiological foundations of learning can be pursued most profitably by measuring physical and chemical changes that influence the transmission of nerve impulses. It has long been established that chemical changes are part of the process of neural transmission; and it is widely agreed that, in some way, biochemical activities also are responsible for all forms of learning, including transfer of training.

One popular theory in the 1960s was that learning and remembering depend on changes in the molecular structure of such chemicals as ribonucleic acid (RNA) and peptides that are incorporated in the cells of the body, including nerve cells. Some researchers have theorized that memory traces are physically coded within the molecules of cells.

Reports of experiments have been published offering evidence that skills have been transferred from one individual to another by injecting materials taken from the brains (or even other parts of the body) of trained animals into the bodies of untrained organisms (*e.g.*, flatworms, rats, hamsters). These reports have encouraged many to hope that someday one might be able to learn a foreign language, for example, by simply taking a pill instead of through the usual time-consuming practice. Subsequent efforts to repeat such experiments sometimes have given positive results but more often have yielded no evidence of chemical transfer of training from one individual to the next. In view of such inconsistent findings, this question became a matter of considerable controversy. Many investigators seemed inclined to dismiss the notion that organisms can learn by swallowing chemicals or through injection as another of those oversimplified interpretations that continue to be offered in efforts to account for complex psychophysiological phenomena. (H.H.K.)

BIBLIOGRAPHY

General works. JOHN F. HALL, *The Psychology of Learning* (1966), provides a comprehensive account of the empirical data that theories of learning are designed to integrate. ERNEST R. HILGARD and GORDON H. BOWER, *Theories of Learning,* 5th ed. (1981), is the standard reference on this subject. ERNEST R. HILGARD and DONALD G. MARQUIS, *Hilgard and Marquis' Conditioning and Learning,* 2nd ed. rev. by GREGORY A. KIMBLE (1961), describes theoretical issues as they apply to simple learning. The separate articles reproduced in GREGORY A. KIMBLE (ed.), *Foundations of Conditioning and Learning* (1967), develop some of these arguments in more detail and also present synopses of the theoretical positions of Hull, Guthrie, and Tolman. GREGORY A. KIMBLE, NORMAN GARMEZY, and EDWARD ZIGLER, *Principles of Psychology,* 6th ed. (1984), is a general textbook that covers several theoretical issues and should be particularly useful on the topics of memory and retrieval. WILLIAM BECHTEL and ADELE ABRAHAMSEN, *Connectionism and the Mind: An Introduction to Parallel Processing in Networks* (1991), gives an overview of connectionist theory and higher-level cognitive processing. DAVID E. RUMELHART *et al., Parallel Distributed Processing: Explorations in the Microstructure of Cognition,* 2 vol. (1986), is a landmark collection of articles on the connectionist theory of learning. PHILIP JOHNSON-LAIRD, *The Computer and the Mind,* 2nd rev. ed. (1993), provides an easy introduction to cognitive science. (G.A.K./Ed.)

Psychomotor learning. A.L. IRION, "A Brief History of Research on the Acquisition of Skill," in EDWARD A. BILODEAU (ed.), *Acquisition of Skill* (1966), contains an excellent historical survey. ROBERT S. WOODWORTH and HAROLD SCHLOSBERG, *Woodworth & Schlosberg's Experimental Psychology,* 3rd ed. by J.W. KLING *et al.,* 2 vol. (1972), is one of the best standard reference works. CLYDE E. NOBLE, "S-O-R and the Psychology of Human Learning," *Psychological Reports,* 18:923–943 (1966), offers a brief introduction to human learning. ARTHUR WEEVER MELTON (ed.), *Apparatus Tests* (1947), is a classic volume on psychomotor devices used in aviation psychology. Reviews of the psychomotor field can be found in two articles in *Annual Review of Psychology:* E.A. BILODEAU and I.MCD. BILODEAU, "Motor-Skills Learning," 12:243–280 (1961); and CLYDE E. NOBLE, "The Learning of Psychomotor Skills," 19:203–250 (1968); and in J.A. ADAMS, "Historical Review and Appraisal of Research on the Learning, Retention, and Transfer of Human Motor Skills," *Psychological Bulletin,* 101(1):41–47 (January 1987). CLARK L. HULL, *Principles of Behavior* (1943, reissued 1966); and KENNETH W. SPENCE, *Behavior Theory and Conditioning* (1956, reprinted 1978), are two of the most influential books on general behaviour theory from the reinforcement viewpoint. CLYDE E. NOBLE, "A Theory of Psychomotor Skill: Derivation and Data," *Psychonomic Science,* 21:344 (1970), makes a specific application to psychomotor learning. RICHARD A. SCHMIDT, *Motor Control and Learning: A Behavioral Emphasis,* 2nd ed. (1988), discusses motor learning from a behavioral and physiological perspective. Two volumes presenting detailed information-processing analyses of skill are PAUL M. FITTS and MICHAEL I. POSNER, *Human Performance* (1967, reprinted 1979); and A.T. WELFORD, *Fundamentals of Skill* (1968). K.M. NEWELL, "Motor Skill Acquisition," *Annual Review of Psychology,* 42:213–237 (1991), presents a more recent discussion of information processing in the acquisition of motor skills. E.A. BILODEAU (ed.), *Principles of Skill Acquisition* (1969), provides an eclectic, simplified treatment of current topics by several authors. ROBERT N. SINGER, *Motor Learning and Human Performance,* 3rd ed. (1980), is oriented mainly toward athletic proficiency and physical education. Articles of specialized interest, as indicated by their titles, are J.A. ADAMS, "Response Feedback and Learning," *Psychological Bulletin,* 70:486–504 (1968); CLYDE E. NOBLE, "Acquisition of Pursuit Tracking Skill Under Extended Training As a Joint Function of Sex and Initial Ability," *Journal of Experimental Psychology,* 86:360–373 (December 1970); and R.B. PAYNE, "Functional Properties of Supplementary Feedback Stimuli," *Journal of Motor Behavior,* 2:37–43 (1970). (C.E.N./Ed.)

Concept formation. LYLE E. BOURNE, JR., *Human Conceptual Behavior* (1966), provides a well-organized review of laboratory studies. JEROME S. BRUNER, JACQUELINE J. GOODNOW, and GEORGE A. AUSTIN, *A Study of Thinking* (1956, reprinted 1986), is a classic, well-written description of some experiments in the field. JOHN H. FLAVELL, *The Developmental Psychology of Jean Piaget* (1963), offers a competent explanation of Piaget's controversial and provocative ideas. EARL B. HUNT, *Concept Learning: An Information Processing Problem* (1962, reprinted 1974), gives a comprehensive review and theoretical presentation of major approaches. B.F. SKINNER, *Verbal Behavior* (1957), makes a challenging proposal for applying laboratory psychology. However, a critical review is presented by NOAM CHOMSKY, "A Review of B.F. Skinner's *Verbal Behavior,*" in JERRY A. FODOR and JERROLD J. KATZ (eds.), *The Structure of Language: Readings in the Philosophy of Language* (1964), pp. 547–578. Works considering specialized topics include STEVEN PINKER, *Language Learnability and Language Development* (1984), which extends theories of language acquisition; and HENRY M. WELLMAN, *The Child's Theory of the Mind* (1990). (E.B.H./Ed.)

Transfer of training. Introductory psychology texts that discuss transfer of training as an experimental phenomenon and its implications for a wide variety of behaviour include HOWARD H. KENDLER, *Basic Psychology,* 3rd ed. (1974); and HOWARD H. KENDLER and TRACY S. KENDLER, *Basic Psychology: Brief Edition* (1971). Undergraduate texts that review theories and experimental evidence concerning transfer of training are JAMES DEESE and STEWART H. HULSE, *The Psychology of Learning,* 5th

ed. (1980); and JOHN F. HALL, *The Psychology of Learning* (1966). HENRY C. ELLIS, *The Transfer of Learning* (1965), presents a general analysis of transfer and includes reprints of important journal articles on the topic. HAYNE W. REESE, *The Perception of Stimulus Relations: Discrimination Learning and Transposition* (1968); and DONALD A. RILEY, *Discrimination Learning* (1968), are accounts of how transfer of training influences discrimination learning. More recent discussion of research on transfer of training may be found in STEVEN M. CORMIER and JOSEPH D. HAGMAN (eds.), *Transfer of Learning: Contemporary Research and Applications* (1987).

(H.H.K./Ed.)

Lebanon

The Republic of Lebanon (al-Jumhūrīyah al-Lubnānīyah), a predominantly mountainous country of great scenic beauty, is an Arab republic situated on the eastern shore of the Mediterranean Sea. Consisting of a narrow strip of territory approximately 135 miles (215 kilometres) long from north to south and 20 to 55 miles wide from east to west, the country is bounded to the north and east by Syria and to the south by Israel. With an area of 3,950 square miles (10,230 square kilometres), Lebanon is one of the world's smaller sovereign states. The capital is Beirut.

Though Lebanon, particularly its coastal region, was the site of some of the oldest human settlements in the world—the Phoenician ports of Tyre (modern Ṣūr), Sidon (Ṣaydā), and Byblos (Jubayl) were dominant centres of trade and culture in the 3rd millennium BC—it was not until 1920 that the contemporary state came into being. In that year France, which administered Lebanon as a League of Nations mandate, established the state of Greater Lebanon. Lebanon then became a republic in 1926 and achieved independence in 1943.

As an Arab republic, Lebanon shares many of the cultural characteristics of the Arab world, yet it has attributes that differentiate it from many of its Arab neighbours. Its rugged, mountainous terrain has served throughout history as an asylum for diverse religious and ethnic groups and for political dissidents. Lebanon is one of the most densely populated countries in the Mediterranean area. It has one of the highest rates of literacy. Although its prosperity is unevenly distributed, having bypassed large segments of its population, wealth and privilege appear to be evenly distributed among its middle-income group. Notwithstanding its meagre natural resources, Lebanon long managed to serve as a busy commercial and cultural centre for the Middle East.

This outward image of vitality and growth nevertheless disguised serious problems. Not only did Lebanon have to grapple with internal problems of social and economic organization, but also it had to struggle to define its position in relation to Israel, to its Arab neighbours, and to Palestinian refugees living in Lebanon. The Lebanese pluralistic communal structure eventually collapsed under the pressures of this struggle. Communal rivalries over political power became so exacerbated by the complex issues that arose from the Palestinian question that a breakdown of the governmental system resulted from an extremely damaging civil war that began in 1975.

This article is divided into the following sections:

Physical and human geography

THE LAND

Relief. As in any mountainous region, the physical geography of Lebanon is extremely complex and varied. Landforms, climate, soils, and vegetation undergo some sharp and striking changes within short distances. Four distinct physiographic regions may be distinguished: a narrow coastal plain along the Mediterranean Sea, the Lebanon Mountains (Jabal Lubnān), al-Biqāʿ (Beqaa) valley, and the Anti-Lebanon and Hermon ranges running parallel to the coastal mountains.

Physiographic regions

The coastal plain is narrow and discontinuous, almost disappearing in places. It is formed of river-deposited alluvium and marine sediments, which alternate suddenly with rocky beaches and sandy bays, and is generally fertile. In the far north it expands to form the ʿAkkār Plain.

The snowcapped Lebanon Mountains are the most prominent feature of the country's landscape. The range, rising steeply from the coast, forms a ridge of limestone and sandstone, cut by narrow and deep gorges. It is approximately 100 miles long and varies in width from 35 to six miles. Its maximum elevation is at Qurnat as-Sawdāʾ (10,138 feet [3,090 metres]) in the north, where the renowned cedars of Lebanon grow in the shadow of the peak. The range then gradually slopes to the south, rising again to a second peak, Jabal Ṣannīn, northeast of Beirut. To the south the range gives way to the hills of Galilee, which are lower. The limestone composition of the mountains provides a relatively poor topsoil. The lower and middle slopes, however, are intensively cultivated, the terraced hills standing as a scenic relic of the ingenious tillers of the past. On the coast and in the northern mountains reddish topsoils with a high clay content retain moisture and provide fertile land for agriculture, although they are subject to considerable erosion.

Al-Biqāʿ valley lies between the Lebanon Mountains in the west and the Anti-Lebanon Mountains in the east; its fertile soils consist of alluvial deposits from the mountains on either side. The valley, approximately 110 miles long

Al-Biqāʿ

MAP INDEX

Political subdivisions
Bayrūt ....33 56 N 35 30 E
Biqā', al- ....34 00 N 36 10 E
Jabal Lubnān ....33 50 N 35 40 E
Janūb, al- ....33 20 N 35 18 E
Nabaṭiyah, an- ....33 20 N 35 35 E
Shamāl, ash- .....34 25 N 36 00 E

Cities and towns
'Abdah, al- ....34 31 N 35 58 E
Amyūn ....34 18 N 35 49 E
B'abdā ....33 50 N 35 32 E
Ba'labakk ....34 00 N 36 12 E
B'aqlīn ....33 41 N 35 33 E
Batrūn, al- ....34 15 N 35 39 E
Beirut (Bayrūt) ....33 53 N 35 30 E
Bḥamdūn ....33 48 N 35 39 E
Bint Jubayl ......33 07 N 35 26 E
Bsharrī ....34 15 N 36 01 E
Byblos, see Jubayl
Dāmūr, ad- ......33 44 N 35 27 E
Ghazīr ....34 01 N 35 40 E
Ghazzah ....33 40 N 35 49 E
Ghūmā ....34 13 N 35 42 E
Ḥalbā ....34 33 N 36 05 E
Ḥaṣbayya ....33 24 N 35 41 E
Ḥimlāyā ....33 56 N 35 42 E
Hirmil, al- ....34 23 N 36 23 E
Ihdin ....34 17 N 35 58 E
Jazzīn ....33 32 N 35 34 E
Jbā' ....33 29 N 35 31 E
Jubayl (Byblos) ...34 07 N 35 39 E
Jubb Jannīn .....33 37 N 35 47 E
Jūniyah ....33 59 N 35 38 E
Jwayyā ....33 14 N 35 19 E
Khaldah ....33 47 N 35 29 E
Labwah, al- ....34 12 N 36 21 E
Marj 'Uyūn ....33 22 N 35 35 E
Mashgharah .....33 32 N 35 39 E
Mīnā', al- ....34 27 N 35 49 E
Nabaṭīyah at-Taḥtā, an- ....33 23 N 35 29 E
Nāqūrah, an- .....33 07 N 35 08 E
Qarṭabā ....34 06 N 35 51 E
Qubayyāt, al- .....34 34 N 36 17 E
Rāshayyā ....33 30 N 35 51 E
Ṣarafand, aṣ- .....33 27 N 35 18 E
Ṣayda, see Sidon
Shḥīm ....33 37 N 35 29 E
Shikkā ....34 20 N 35 44 E
Sidon (Ṣayda) ....33 33 N 35 22 E
Sīr aḍ-Ḍinnīyah ...34 23 N 36 02 E
Ṣūr, see Tyre
Tripoli (Ṭarābulus) .....34 26 N 35 51 E
Tyre (Ṣūr) ....33 16 N 35 11 E
Zaḥlah ....33 51 N 35 53 E
Zghartā ....34 24 N 35 54 E

Physical features and points of interest
Abū 'Alī, *river* .....34 27 N 35 50 E
'Akkār, Sahl, *plain* ....34 39 N 36 03 E
'Anjar, *archaeological site* .....33 43 N 35 57 E
Anti-Lebanon (Jabal ash-Sharqī, al-), *mountains* ......34 00 N 36 30 E
'Arabah, Jabal al-, *mountains* ...33 35 N 35 46 E
'Āṣī, al-, see Orontes
Awwalī, al-, *river* ..33 35 N 35 23 E
Ba'labakk, *archaeological site* .....34 01 N 36 11 E
Bārūk, Jabal al-, *mountain* ....33 43 N 35 44 E
Bayrūt, *river* .....33 54 N 35 32 E
Biqā', al-, *valley* ...33 50 N 36 00 E
Byblos, *archaeological site* .....34 07 N 35 38 E
Cedars of Lebanon, *cedar grove* ....33 41 N 35 41 E
Dāmūr, ad-, *river* ..33 42 N 35 26 E
Ḥāṣbānī, al-, *river* ..33 11 N 35 37 E
Hermon, Mt. (Shaykh, Jabal ash-), *mountain* ..33 26 N 35 51 E
Ibrāhīm, *river* .....34 04 N 35 38 E
Ithnayn, Jabal al-, *mountain* ....34 03 N 36 29 E
'Izzīyah, al-, *river* ..33 12 N 35 12 E
Jabal ash-Sharqī, al-, see Anti-Lebanon
Jawz, al-, *river* ....34 16 N 35 39 E
Kabīr, al-, *river* ....34 38 N 35 58 E
Kalb, al-, *river* .....33 57 N 35 35 E
Kanīsah, Jabal al-, *ridge* ....33 50 N 35 47 E
Lebanon Mountains (Lubnān, Jabal) ..34 00 N 36 00 E
Līṭānī, al-, *river* ....33 20 N 35 14 E
Mediterranean Sea ....34 00 N 35 10 E
Mūsā, al-, *river* ....34 26 N 36 00 E
Orontes ('Āṣī, al-), *river* ....34 15 N 36 20 E
Qir'awn, Buḥayrat al-, *reservoir* ....33 34 N 35 42 E
Salūqī, as-, *watercourse* .....33 18 N 35 26 E
Ṣannīn, Jabal, *mountain* ....33 57 N 35 52 E
Sawdā', Qurnat as-, *mountain* ...34 18 N 36 07 E
Shamīs al-Ḥijār, *mountain* ....34 04 N 36 01 E
Sharbīn, ash-, *watercourse* .....34 27 N 36 26 E
Shaykh, Jabal ash-, see Hermon, Mt.
Ṣūr, see Tyre
Tyre (Ṣūr), *archaeological site* .....33 16 N 35 11 E
Wadīn, *watercourse* .....34 35 N 36 19 E
Zahrānī, az-, *river* .33 29 N 35 20 E

LEBANON

and from six to 16 miles wide, is part of the great East African Rift System. In the south, al-Biqā' becomes hilly and rugged, blending into the foothills of Mt. Hermon (Jabal ash-Shaykh) to form the upper Jordan Valley.

The Anti-Lebanon range (al-Jabal ash-Sharqī) starts with a high peak in the north and slopes southward until it is interrupted by Mt. Hermon (9,232 feet).

Drainage. Lebanese rivers, though numerous, are mostly winter torrents, draining the western slopes of the Lebanon Mountains. The only exception is the Līṭānī (90 miles long), which rises near the famed ruins of Baalbek (Ba'labakk) and flows southward in al-Biqā' to empty into the Mediterranean near historic Tyre. The two other important rivers are the Orontes (Nahr al-'Āṣī), which rises in the north of al-Biqā' and flows northward, and the Kabīr.

Climate. There are sharp local contrasts in climatic conditions. Lebanon is included in the Mediterranean climatic region, which extends westward to the Atlantic Ocean. The winter storms formed over the ocean move eastward through the Mediterranean, bringing rain at that season; in summer the Mediterranean receives no rain. The climate of Lebanon is generally subtropical and is characterized by hot, dry summers and mild, humid winters. Mean daily maximum temperatures on the coast and in al-Biqā' range from 90° F (32° C) in July to 60° F (16° C) on the coast and 50° F (10° C) in al-Biqā' in January. Mean minimum temperatures in January are 50° F (10° C) on the coast and 35° F (2° C) in al-Biqā'. At 5,000 feet, the altitude of the highest settlements, these are reduced by about 15° F (8° C).

Nearly all precipitation falls in winter and averages 30 to 40 inches (750 to 1,000 millimetres) on the coast, rising to more than 50 inches in higher altitudes. Al-Biqā' is drier and receives 15 to 25 inches. On the higher mountaintops, this precipitation falls as heavy snow that remains until early summer.

Cedar trees

Plant and animal life. Lebanon was heavily forested in ancient and medieval times, and its timber—particularly its famed cedar—was exported for building and shipbuilding. The natural vegetation, however, has been grazed, burned, and cut for so long that little of it is regenerated. What survives is a wild Mediterranean vegetation of brush and low trees, mostly oaks, pines, cypresses, firs, junipers, and carobs.

Few large wild animals survive in Lebanon, though bears are occasionally seen in the mountains. Among the smaller animals, deer, wildcats, hedgehogs, squirrels, martens, dormice, and hares are found. Numerous migratory birds from Africa and Europe visit Lebanon. Flamingos, pelicans, cormorants, ducks, herons, and snipes frequent the marshes; eagles, buzzards, kites, falcons, and hawks inhabit the mountains; and owls, kingfishers, cuckoos, and woodpeckers are common.

Settlement patterns. Most of the population live on the coastal plain, and progressively fewer people are found farther inland. Rural villages are sited according to water supply and the availability of land, frequently including terraced agriculture in the mountains. Northern villages are relatively prosperous and have some modern architecture. Villages in the south have been generally poorer and less stable; their agricultural land is less fertile and, because of their proximity to Israel, many have been subject to frequent dislocation, invasion, and destruction since 1975. Most cities are located on the coast; they have been inundated by migrants and displaced persons, and numerous, often poor, suburbs have been created as a result. Before 1975 many villages and cities were composed of several different religious groups, usually living together in harmony, and rural architecture reflected a unity of style irrespective of religious identity. Since the civil war began, a realignment has moved thousands of Christians north of Beirut along the coast and thousands of Muslims south or east of Beirut, so that settlement patterns reflect the chasms separating sections of the Lebanese people from each other.

THE PEOPLE

Lebanon has a heterogeneous society composed of numerous ethnic, religious, and kinship groups. Primordial attachments and local communalism antedate the creation of the present territorial and political entity and continue to survive with remarkable tenacity.

Ethnic and linguistic groups. Ethnically, the Lebanese compose a mixture in which Phoenician, Greek, Armenian, and Arab elements are discernible. Arabic is the official language, but French and English are widely spoken. A small percentage of the population is Armenian-speaking, and Syriac is used in some of the churches of the Maronites (Roman Catholics following an Eastern rite).

Religious groups. Perhaps the most distinctive feature of Lebanon's social structure is its varied religious composition. Since the 7th century Lebanon has served as a refuge for persecuted Christian and Muslim sects. The population is estimated to consist of a majority of Muslims and a large minority of Christians. Shī'ite Muslims are the most numerous group. Among the Christians, Maronites form the largest group, and Greek Orthodox and Greek Catholics are the next largest groups. Among the three Muslim denominations, the Shī'ites are followed closely by the Sunnites; the Druzes constitute a small percentage. There is also a very small minority of Jews.

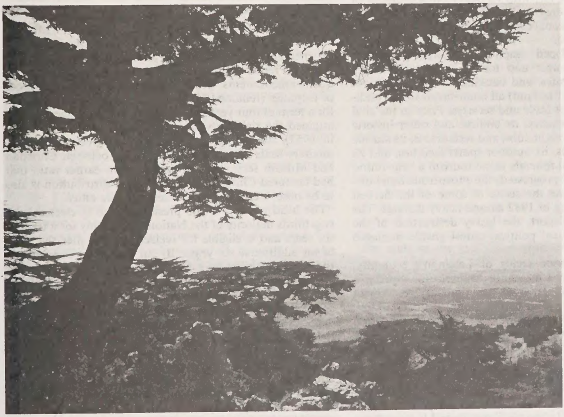
Demography. One of the most salient demographic features of Lebanon is the uneven distribution of its population. The country's overall density is much lower than that of Bayrūt *muḥāfaẓah* (Beirut governorate) but much higher than that of the most sparsely populated, al-Biqā' governorate.

Before the civil war began, the movement of people from rural areas was a major factor in the country's soaring rate of urbanization. Most of the internal migration was to Beirut, which accounted for the great majority of Lebanon's urban population. The civil war and postwar fighting led to a substantial return of people to their villages and to a large migration abroad, primarily to the United States, Europe, Latin America, Australia, and the oil states of the Middle East.

THE ECONOMY

Until 1975 Lebanon had an economy characterized by a minimum of government intervention in private enterprise. Since the civil war, the weak central government has exercised little power in economic matters, and local militias have dominated public decision making.

Werner Braun—Camera Press



Cedars of Lebanon on the slopes of the Lebanon Mountains.

The services sector generated the overwhelming proportion of national income before the civil war and employed the largest proportion of the labour force; industry generated the second largest proportion of income and of employment. Agriculture accounted for a smaller proportion of income. The growth of services was related mainly to international transport and trade and to the position of Beirut as a centre of international banking and tourism.

Economic losses after 1974

The war of 1975–76, the Israeli invasion of 1982, and the continuing violence have left deep scars and have led to chaos in the economy. There has been extensive destruction in all sectors, but especially in housing, trade, and public services, and the country's productive capacity has been drastically reduced. The greatest reduction in productive capacity seems to be in services, followed by industry and agriculture.

Resources. The mineral resources of Lebanon are few. There are deposits of high-grade iron ore and lignite; building-stone quarries; high-quality sand, suitable for glass manufacture; and lime. The Liṭānī River hydroelectric project generates electricity and also has increased the amount of irrigated land for agriculture.

Agriculture. Arable land is scarce, but the climate and the relatively abundant water supply from springs favour the intensive cultivation of a variety of crops on mountain slopes and in the coastal region. On the irrigated coastal plain, market vegetables, bananas, and citrus crops are grown. In the foothills the principal crops are olives, grapes, tobacco, figs, and almonds. At higher altitudes (about 1,500 feet) peaches, apricots, plums, and cherries are planted, while apples and pears thrive at an altitude of about 3,000 feet. Sugar beets, cereals, and vegetables are the main crops cultivated in al-Biqāʿ. Poultry is a major source of agricultural income, and goats, sheep, and cattle are also raised.

As a result of the continued violence, many small farmers have lost their livestock, and there has been a noticeable decrease in the production of many agricultural crops. The production of hemp, the source of hashish, has flourished in al-Biqāʿ valley, however, and the hashish is exported illegally through ports along the coast.

Industry. The majority of the country's industry survived the civil war unscathed. Beirut's industrial belt was razed, but some of the country's large complexes were unharmed. Manufacturing recovered to more than half of the still-existing capacity, restrained by limited labour mobility, difficulty in acquiring supplies, insufficient working capital, and difficulty in obtaining credit. The Israeli invasion of 1982, however, with its heavy bombardment of some of Lebanon's major cities and subsequent sabotage by local warring factions, caused further damage to industry and infrastructure.

Beirut's well-developed seaport and airport and the country's free economic and foreign-exchange systems, favourable interest rates, and banking secrecy law (modeled upon that of Switzerland) all contributed to the traditional preeminence of trade and services. Prior to the civil war, the country's scenery, its biblical and other historic sites, its hotels, bars, nightclubs, and restaurants, its seaside and mountain resorts, its outdoor sports facilities, and its international cultural festivals made tourism a year-round industry. As the war progressed, the prosperous hotel district in Beirut became the scene of some of the fiercest fighting, and bombing in 1982 caused heavy damage. The closures of Beirut airport, the heavy destruction of the port, and the continued political unrest greatly damaged the service industry.

Finance. The finance sector of Lebanon's economy, including banking and insurance, showed an impressive expansion before the war, and the monetary reserves of Lebanon continued to rise despite political uncertainties. During the two years of civil war and the extended period of domestic instability and economic inflation following that, reserves, which included a considerable portion of gold, nevertheless continued to rise.

Capital inflows

The balance of payments has traditionally shown a surplus. The strength of the Lebanese pound and of the balance-of-payments position reflected large inflows of capital, mostly from Lebanese living abroad (whose numbers rose considerably during and after the civil war) and from the high level of liquidity of commercial banks. By 1983, however, inflows from Lebanese living abroad had begun to decrease, and the value of the Lebanese pound fell dramatically.

Trade. Widespread smuggling, covert foreign aid to armed groups, and illegal drug production have disguised the pattern of trade since 1975. Exports, chiefly vegetable products, textiles, and nonprecious metals are sent mainly to Middle Eastern countries. Imports such as consumer goods, machinery and transport equipment, petroleum products, and food come mostly from western Europe. A huge trade deficit has been partly covered by "invisible" items such as foreign remittances and government loans.

Trade unions. Lebanon is one of the few countries in the Middle East with a comparatively well-developed labour movement. Trade unions have secured some tangible gains, such as fringe benefits, collective bargaining contracts, and better working conditions. During the civil war, divisions in many of the trade unions weakened their normal functions, and many of their members joined the warring factions. Many others emigrated.

Transportation. As in antiquity, Lebanon's situation makes it a vital crossroads between East and West. The road network traversing Lebanon includes international highways, which form part of major land routes connecting Europe with the Arab countries and the East. There are also national highways, paved secondary roads, and unpaved roads. The railway system, which includes lines along the coast and up al-Biqāʿ valley and a cog railway across the Lebanon Mountains, connects with the rail system of Syria. In the past the system was linked with rail systems of other Arab countries and with Europe and was used mainly for long-distance bulk transport.

Numerous ports lie along the seacoast. Berths for oil tankers have been built offshore at Tripoli and at az-Zahrānī, near Sidon, where pipeline terminals and refineries also are located. The principal cargo and passenger port is that of Beirut, which has a free zone and storage facilities for transit shipments. The port has been expanded and deepened, and a large storage silo (for wheat and other grains) has been built, but port facilities were severely damaged during the civil war and the postwar fighting. The harbour at Jūniyah has grown in importance.

The port of Beirut

Beirut International Airport was one of the busiest airports in the Middle East before the civil war. Its runways were built to handle the largest jet airplanes in service, and a number of international airlines used Beirut regularly.

GOVERNMENT AND SOCIAL CONDITIONS

Government. *The constitutional framework.* Modern Lebanon is a republic with a parliamentary system of government. Its constitution, promulgated in 1926 during the French mandate and modified by several subsequent amendments, provides for a unicameral Chamber of Deputies (renamed National Assembly in 1979) elected for a term of four years by universal adult suffrage (women attained the right to vote and eligibility to run for office in 1953). According to the 1989 Ṭā'if agreement, parliamentary seats are apportioned equally between Christian and Muslim sects, thereby replacing an earlier ratio that had favoured Christians. This sectarian distribution is also to be observed in appointments to public office.

The head of state is the president, who is elected by a two-thirds majority of the National Assembly for a term of six years and is eligible for reelection only after the lapse of an additional six years. By an unwritten convention, the president must be a Maronite Christian, the premier a Sunnite Muslim, and the speaker of the National Assembly a Shīʿite. The president, in consultation with the speaker of the National Assembly and the parliamentary deputies, invites a Sunnite Muslim to form a Cabinet, and the Cabinet members' portfolios are organized to reflect the sectarian balance. The Cabinet, which holds more executive power than the president, requires a vote of confidence from the assembly in order to remain in power. A vote of no confidence, however, is rarely exercised in practice. A Cabinet usually falls because of internal dissension, societal strife, or pressure exerted by foreign states.

Local government. Lebanon is divided into *muḥāfaẓāt* (governorates): Bayrūt (Beirut), Jabal Lubnān, ash-Shamāl, al-Janūb, al-Biqāʿ, and an-Nabaṭīyah. These are administered by the *muḥāfiẓ* (governor), who represents the central government. The *muḥāfaẓāt* are further divided into *qadawat* (districts), each of which is presided over by a *qāʾim-maqām* (district chief), who, along with the *muḥāfiẓ*, supervises local government. Municipalities (communities with at least 500 inhabitants) elect their own councils, which in turn elect mayors and vice mayors. Villages and towns (more than 50 and fewer than 500 inhabitants) elect a *mukhṭār* (headman) and a council of elders, who serve on an honorary basis. All officers of local governments serve four-year terms.

The political process. The political system in Lebanon remains a curious blend of secular and traditional features. Until 1975 the country appeared to support liberal and democratic institutions, yet, in effect, it had hardly any of the political instruments of a civil polity. Its political parties, parliamentary blocs, and pressure groups were so closely identified with parochial, communal, and personal loyalties that they often failed to serve the larger national purpose of the society. The National Pact of 1943, a sort of Christian-Muslim entente, sustained the national entity (*al-kiyān*), yet this sense of identity was neither national nor civic.

In April 1975 the political process collapsed. The war that had engulfed the Lebanese exposed the vulnerability of the political system. The legitimate authority continues to maintain the facade of continuity, while the process on which it is based was destroyed by the contending forces in the conflict that has continued to ravage Lebanon. The control of the official central government is precarious; sectarian militias and foreign countries exert great influence.

Justice. The system of law and justice is mostly modeled on French concepts. The judiciary consists of courts of the first instance, courts of appeal, courts of cassation, and a Court of Justice that handles cases affecting state security. The Council of State is a court that deals with administrative affairs. In addition, there are religious courts that deal with matters of personal status (such as inheritance, marriage, and property matters) as they pertain to autonomous communities. Despite the country's well-developed legal system and a very high proportion of lawyers, significant numbers of disputes and personal grievances are resolved outside the courts. Justice by feud and vendetta continues.

Armed forces. The armed forces consist of an army, air force, and navy. Lebanon also has a paramilitary gendarmerie and police force. During the civil war the army practically disintegrated as splinter groups joined the different warring factions. Reconstruction of the Lebanese armed forces has been attempted, particularly with the assistance first of the United States and then of Syria, with substantial effect. Responsibility for maintaining security and order has often fallen to the various political and religious factions and to foreign occupiers.

Education. A well-developed system of education reaches all levels of the population. Literacy is among the highest in the Middle East. Education was once almost exclusively the responsibility of religious communities or foreign groups, but the number of students in public schools has risen to about two-fifths of the total school enrollment.

The compulsory five-year primary school program is followed either by a seven-year secondary program (leading to the official baccalaureate certificate) or by a four-year program of technical or vocational training. Major universities include the American University of Beirut, the Université Saint-Joseph (subsidized by the French government and administered by the Jesuit order), the Lebanese University (Université Libanaise), and the Beirut Arab University (an affiliate of the University of Alexandria).

Universities

Health and welfare. Public health services are largely concentrated in the cities, although the government increasingly directs medical aid into rural areas. As in the field of social welfare, nongovernmental voluntary associations—mostly religious, communal, or ethnic—are active. The Lebanese diet is generally satisfactory, and the high standard of living and the favourable climate have served to reduce the incidence of many diseases that are still common in other Middle Eastern countries.

Lebanon has a large number of skilled medical personnel, and hospital facilities are also adequate under normal circumstances. The thousands of casualties and deaths caused by almost constant warfare since 1975 have overburdened the country's health and medical facilities and hampered the treatment of routine illnesses.

The National Social Security Fund, which is not fully implemented, provides sickness and maternity insurance, labour accident and occupational disease insurance, family benefits, and termination-of-service benefits.

Housing. In response to the need for low-cost housing, the Popular Housing Law was enacted, providing for the rehabilitation of substandard housing. Prior to the civil war a substantial percentage of homes were without bathrooms, and thousands of families, including Palestinian refugees, were living in improvised accommodations. When an economic boom attracted villagers to the capital, the housing shortage worsened considerably. The civil war drastically increased the problem. Thousands of homes in battle zones were destroyed, and entire villages were evacuated and others occupied. The result was chaos in which property rights were violated as a matter of course. The government, in an attempt to remedy the situation, set up a Housing Bank to make housing loans.

Housing Bank

Wages and cost of living. A minimum wage is set by the Labour Code, and legislation provides for cost-of-living increases. The cost of living increased sharply prior to, during, and after the civil war, mainly because of a substantial rise in the cost of rent, education, food, and petroleum products.

Social and economic division. Lebanese society was able for a long time to give a semblance of relative economic stability. The existence of a large middle-income group, in addition to the political and social legitimacy of kinship ties and religious and communal attachments, reinforced the veneer that masked the growing socioeconomic dislocations. The interaction of these factors covered up the growing class polarization, especially around the industrial belt that encircled Beirut. The eruption of civil conflict in 1975 and the ensuing state of chaos is attributable in part to the fact that the system of government was unresponsive to the acute social problems and grievances.

CULTURAL LIFE

The cultural milieu. Historically, Lebanon is the heir of a long succession of Mediterranean cultures—Phoenician, Greek, and Arab. Its cultural milieu continues to show clear manifestations of a rich and diverse heritage. As an Arab country, Lebanon shares more than a common language with neighbouring Arab states; it also has a similar cultural heritage and common interests.

In the 19th century Lebanese linguists were in the vanguard of the Arabic literary awakening. In more recent times, writers of the calibre of Khalil Gibran, Georges Shehade, and Michel Chiha have been widely translated and have reached an international audience.

While for a time cultural life in Lebanon was predominantly centred around universities and affiliated institutions, there has been an impressive proliferation of cultural activities under other auspices. Beirut has several museums and a number of private libraries, learned societies, and research institutions.

The state of the arts. Lebanon's antiquities and ruins have provided not only inspiration for artists but also magnificent backdrops for annual music festivals, most notably the Baalbek International Festival. At one time, international opera, ballet, symphony, and drama companies, of nearly all nationalities, competed to enrich the cultural life of Beirut. Lebanon has produced a number of gifted young artists who have shown a refreshing readiness to experiment with new expressive forms. Some Lebanese are active in European opera and theatre companies, while others are intent on creating a wider audience for classical Arabic music and theatre.

The cultural awakening encouraged the revival of national folk arts, particularly song, *dabkah* (the national dance), and *zajal* (folk poetry), and the refinement of traditional crafts. Although the Baalbek International Festival was suspended during the civil war, popular theatre and radio satires continued to flourish in the war-ridden country.

The communications media. In addition to the wide variety of foreign newspapers and magazines that can be found in Beirut, Lebanon has registered publications in Arabic, English, French, and Armenian, including a number of daily newspapers.

Because of the wide choice of films, comfortable surroundings, and low prices, going to view motion pictures emerged as a popular form of entertainment among the Lebanese. Moviegoing has been supplanted to some extent, however, by private viewing of videocassettes.

Television programming is offered by Beirut's private companies, and Egyptian and Syrian broadcasts are received. With the advent of the transistor, Lebanon became virtually saturated with radios. The government-run radio station broadcasts Arabic, French, English, and Armenian programs, and clandestine radio stations broadcast the news and views of the warring parties.

For statistical data on the land and people of Lebanon, see the *Britannica World Data* section in the BRITANNICA BOOK OF THE YEAR. (S.G.K./C.F.M./W.L.O.)

History

PHOENICIA

Origins and relations with Egypt. The evidence of tools found in caves along the coast of Lebanon shows that the area was inhabited from the Paleolithic through the Neolithic periods. Village life followed the domestication of plants and animals (the Neolithic Revolution, after about 10,000 BC), with Byblos (modern Jubayl) apparently taking the lead. At this site also appear the first traces in Lebanon of pottery and metallurgy (first copper, then bronze, an alloy of tin and copper) by the 4th millennium BC. The Phoenicians, indistinguishable from the Canaanites of Palestine, probably arrived in the land that became Phoenicia (a Greek term applied to the coast of Lebanon) in about 3000 BC. Herodotus and other Classical writers preserve a tradition that they came from the coast of the Erythraean Sea (*i.e.*, the Persian Gulf), but in fact nothing certain is known of their original homeland.

Arrival of the Phoenicians

Except at Byblos, no excavations have produced any information concerning the 3rd millennium in Phoenicia before the advent of the Phoenicians. At Byblos, the first urban settlement is dated *c.* 3050–2850 BC. Commercial and religious connections, probably by sea, with Egypt are attested from the Egyptian 4th dynasty (*c.* 2613–*c.* 2494 BC). The earliest artistic representations of Phoenicians are found in a damaged relief at Memphis of Pharaoh Sahure of the 5th dynasty (early 25th century BC). This shows the arrival of an Asiatic princess to be the Pharaoh's bride; her escort is a fleet of seagoing ships, probably of the type known to the Egyptians as "Byblos ships," manned by crews of Asiatics, evidently Phoenicians.

Byblos was destroyed by fire about 2150 BC, probably by the invading Amorites. The Amorites rebuilt on the site, and a period of close contact with Egypt was begun. Costly gifts were given by the pharaohs to those Phoenician and Syrian princes, such as the rulers of Ugarit and Katna, who were loyal to Egypt. Whether this attests to Egypt's political dominion over Phoenicia at this time or simply to strong diplomatic and commercial relations is not entirely clear.

In the 18th century BC new invaders, called Hyksos, destroyed the Amorite rule in Byblos and, passing on to Egypt, brought the Middle Kingdom to an end (*c.* 1720 BC). Little is known about the Hyksos' origin, but they seem to have been ethnically mixed, including a considerable Semitic element, since the Phoenician deities El, Baal, and Anath figured in their pantheon. The rule of the Hyksos in Egypt was brief and their cultural achievement slight, but in this period the links with Phoenicia and Syria were strengthened by the presence of Hyksos aristocracies throughout the region. Pharaoh Ahmose I expelled the Hyksos in about 1567 BC and instituted the New Kingdom policy of conquest in Palestine and Syria. In his annals Ahmose records capturing oxen from the Fenkhw, a term here perhaps referring to the Phoenicians. In the annals of the greatest Egyptian conqueror, Thutmose III (*c.* 1504–1450 BC), the coastal plain of Lebanon, called Djahy, is described as rich with fruit, wine, and grain. Of particular importance to the New Kingdom pharaohs was the timber, notably the cedar, of the Lebanese forests. A temple relief at Karnak depicts the chiefs of Lebanon felling cedars for the Egyptian officers of Seti I (*c.* 1300 BC).

Fuller information about the state of Phoenicia in the 14th century BC comes from the Amarna Letters, diplomatic texts belonging to the Egyptian foreign office, written in cuneiform and found at Tell el-Amarna in Middle Egypt. These archives reveal that the Land of Retenu (Syria–Palestine) was divided into three administrative districts, each under an Egyptian governor. The northernmost district (Amurru) included the coastal region from Ugarit to Byblos; the central (Upi) included the southern al-Biqāʿ valley and Anti-Lebanon; and the third district (Canaan) included all of Palestine from the Egyptian border to Byblos. Also among the letters are many documents addressed by the subject princes of Phoenicia and their Egyptian governors to the pharaoh. It was a time of much political unrest. The Hittites from central Anatolia were invading Syria; nomads from the desert supported the invasion, and many of the local chiefs were ready to seize the opportunity to throw off the yoke of Egypt. The tablets that reveal this state of affairs are written in the language and script of Babylonia (*i.e.*, Akkadian) and thus show the extent to which Babylonian culture had penetrated Palestine and Phoenicia; at the same time they illustrate the closeness of the relations between the Canaanite towns (*i.e.*, those in Palestine) and the dominant power of Egypt.

The Hittite invasion

After the reign of Akhenaton (Amenhotep IV; 1379–*c.* 1362), that power collapsed altogether; but his successors attempted to recover it, and Ramses II reconquered Phoenicia as far as the Nahr el Kelb. In the reign of Ramses III (1198–66) many great changes began to occur as a result of the invasion of Syria by peoples from Asia Minor and Europe. The successors of Ramses III lost their hold over Canaan; the 21st dynasty no longer intervened in the affairs of Syria. In *The Story of Wen-Amon,* a tale of an Egyptian religious functionary sent to Byblos to secure cedar around 1100 BC, the episode of the functionary's inhospitable reception shows the extent of the decline of Egypt's authority in Phoenicia at this time. Sheshonk (Shishak) I, the founder of the 22nd dynasty, in about 928 BC endeavoured to assert the ancient supremacy of Egypt. His successes, however, were not lasting, and, as is clear from the Old Testament, the power of Egypt thereafter became ineffective.

Phoenicia as a colonial and commercial power. Kingship appears to have been the oldest form of Phoenician government. The royal houses claimed divine descent, and the king could not be chosen outside their members. His power, however, was limited by the powerful merchant families, who wielded great influence in public affairs. Associated with the king was a council of elders; such at least was the case at Byblos, Sidon, and perhaps Tyre. During Nebuchadrezzar II's reign (605–562 BC) a republic took the place of the monarchy at Tyre, and the government was administered by a succession of suffetes (judges); they held office for short terms, and in one instance two ruled together for six years. Much later, in the 3rd century BC, an inscription from Tyre also mentions a suffete. Carthage was governed by two suffetes, and these officers are frequently named in connection with the Carthaginian colonies. But this does not justify any inference that Phoenicia itself had such magistrates. Under the Persians a federal bond was formed linking Sidon, Tyre, and Aradus. Federation on a larger scale was never possible in Phoenicia, for the reason that no sense of political unity existed to bind the different states together.

Colonies. By the 2nd millennium BC the Phoenicians had already extended their influence along the coast of the Levant by a series of settlements, some well known, some virtually nothing but names. Well known throughout history are Joppa (Tel Aviv–Yafo) and Dor (later Tantura, modern Nasholim) in the south. The earliest site, however, outside the Phoenician homeland known to possess important aspects of Phoenician culture is Ugarit (Ras Shamra), about six miles north of Latakia. The site was already occupied before the 4th millennium BC, but the Phoenicians only became prominent there in the Egyptian 12th dynasty (1991–1786 BC).

Evidence remains of two temples dedicated to the Phoenician gods Baal and Dagon, although the ruling family appears to have been of different, non-Phoenician stock. The 15th century BC shows strong cultural influences already established there from Cyprus and the world of Mycenaean Greece. A splendid archive of literary and administrative documents found at Ugarit from this period provides evidence of an early form of alphabetic script, arguably the most important Phoenician contribution to Western civilization. In the latter part of the 13th century BC, a flood of land and sea raiders (the Sea Peoples) descended on the Levant coast, destroying many of the Phoenician cities and rolling onward to the frontier of Egypt, from which they were beaten back by the Pharaoh Ramses III. Ugarit was destroyed, together with Aradus and Byblos, though the latter were afterward rebuilt. Though Sidon was destroyed only in part, its inhabitants fled to Tyre, which from this time was regarded as the principal city of Phoenicia and began its period of prosperity and expansion.

Phoenician expansion

Tyre's first colony, Utica in North Africa, was founded perhaps as early as the 10th century BC. It is likely that the expansion of the Phoenicians at the beginning of the 1st millennium BC is to be connected with the alliance of Hiram of Tyre with Solomon of Israel in the second half of the 10th century BC. In the following century, Phoenician presence in the north is shown by inscriptions at Samal (Zincirli Hüyük) in eastern Cilicia, and in the 8th century at Karatepe in the Taurus Mountains, but there is no evidence of direct colonization. Both these cities acted as fortresses commanding the routes through the mountains to the mineral and other wealth of Anatolia.

Cyprus had Phoenician settlements by the 9th century BC. Citium, known to the Greeks as Kition (biblical Kittim), in the southeast corner of the island, became the principal colony of the Phoenicians in Cyprus. Elsewhere in the Mediterranean, several smaller settlements were planted as stepping-stones along the route to Spain and its mineral wealth in silver and copper: at Malta, early remains go back to the 7th century BC, and at Sulcis and Nora in Sardinia and Motya in Sicily, perhaps a century earlier. According to Thucydides, the Phoenicians controlled a large part of the island but withdrew to the northwest corner under pressure from the Greeks. Modern scholars, however, disbelieve this and contend that the Phoenicians arrived only after the Greeks were established.

In North Africa the next site colonized after Utica was Carthage (near Tunis). Carthage in turn seems to have established (or, in some cases, reestablished) a number of settlements in Tunisia, Algeria, Morocco, the Balearic Islands, and southern Spain, eventually making this city the acknowledged leader of the western Phoenicians.

There is little factual evidence to confirm the presence of any settlement in Spain earlier than the 7th century BC, or perhaps the 8th century, and many of these settlements should be viewed as Punic (Carthaginian) rather than Phoenician, though it is likely that the colonizing expeditions of the Carthaginians were supported by many emigrants from the Phoenician homeland. It is very probable that the tremendous colonial activity of the Phoenicians and Carthaginians was stimulated in the 8th to 6th centuries BC by the military blows that were wrecking the trade of the Phoenician homeland in the Levant. Also, competition with the synchronous Greek colonization of the western Mediterranean cannot be ignored as a contributing factor.

In the 3rd century BC Carthage, defeated by the Romans, embarked on a further imperialistic phase in Spain to recoup its losses. Rome responded, defeated Carthage a second time, and annexed Spain. Finally, in 146 BC, after a third war with Rome, Carthage suffered total destruction. It was rebuilt as a Roman colony in 44 BC. The ancient Phoenician language survived in use as a vernacular in some of the smaller cities of North Africa at least until the time of St. Augustine, bishop of Hippo (5th century AD).

Commerce under Egyptian protection

Commerce. The role that tradition especially assigns to the Phoenicians as the merchants of the Levant was first developed on a considerable scale at the time of the Egyptian 18th dynasty. The position of Phoenicia, at a junction of both land and sea routes, under the protection of Egypt, favoured this development, and the discovery of the alphabet and its use and adaptation for commercial purposes assisted the rise of a mercantile society. A fresco in an Egyptian tomb of the 18th dynasty depicted seven Phoenician merchant ships that had just put in at an Egyptian port to sell their goods, including the distinctive Canaanite wine jars in which wine, a drink foreign to the Egyptians, was imported. *The Story of Wen-Amon* recounts the tale of a Phoenician merchant, Werket-el of Tanis in the Nile Delta, who was the owner of "50 ships" that sailed between Tanis and Sidon. The Sidonians are also famous in the poems of Homer as craftsmen, traders, pirates, and slave dealers. The prophet Ezekiel (chapters 27 and 28), in a famous denunciation of the city of Tyre, catalogs the vast extent of its commerce, covering most of the then-known world.

The exports of Phoenicia as a whole included particularly cedar and pine wood from Lebanon, fine linen from Tyre,

Phoenician colonization in the Mediterranean.

Byblos, and Berytos, cloths dyed with the famous Tyrian purple (made from the snail *Murex*), embroideries from Sidon, metalwork and glass, glazed faience, wine, salt, and dried fish. They received in return raw materials, such as papyrus, ivory, ebony, silk, amber, ostrich eggs, spices, incense, horses, gold, silver, copper, iron, tin, jewels, and precious stones.

In addition to these exports and imports, the Phoenicians also conducted an important transit trade, especially in the manufactured goods of Egypt and Babylonia (Herodotus, i, 1). From the lands of the Euphrates and Tigris regular trade routes led to the Mediterranean. In Egypt the Phoenician merchants soon gained a foothold; they alone were able to maintain a profitable trade in the anarchic times of the 22nd and 23rd dynasties (*c.* 945–*c.* 730 BC). Though there were never any regular colonies of Phoenicians in Egypt, the Tyrians had a quarter of their own in Memphis (Herodotus, ii, 112). The Arabian caravan trade in perfume, spices, and incense passed through Phoenician hands on its way to Greece and the West (Herodotus, iii, 107).

Caravan trade

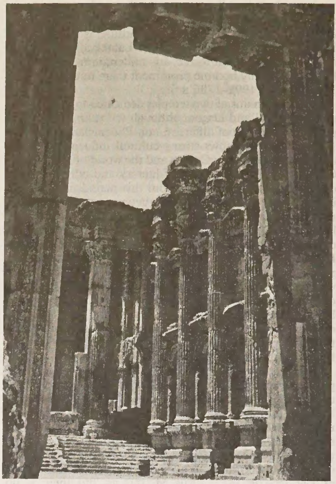
The Phoenicians were not mere passive peddlers in art or commerce. Their achievement in history was a positive contribution, even if it was only that of an intermediary. For example, the extent of the debt of Greece alone to Phoenicia may be fully measured by its adoption, probably in the 8th century BC, of the Phoenician alphabet with very little variation (along with Semitic loan words); by "orientalizing" decorative motifs on pottery and by architectural paradigms; and by the universal use in Greece of the Phoenician standards of weights and measures.

Navigation and seafaring. For the establishment of commercial supremacy, an essential constituent was the Phoenician skill in navigation and seafaring. The Phoenicians are credited with the discovery and use of Polaris (the Pole Star). Fearless and patient navigators, they ventured into regions where no one else dared to go, and always, with an eye to their monopoly, they carefully guarded the secrets of their trade routes and discoveries and their knowledge of winds and currents. Pharaoh Necho II (610–595 BC) organized the Phoenician circumnavigation of Africa (Herodotus, iv, 42). Hanno, a Carthaginian, led another in the mid-5th century. The Carthaginians seem to have reached the island of Corvo in the Azores; and they may even have reached Britain, for many Carthaginian coins have been found there.

Assyrian and Babylonian domination of Phoenicia. Between the withdrawal of Egyptian rule in Syria and the western advance of Assyria there was an interval during which the city-states of Phoenicia owned no suzerain. Byblos had kings of its own, among them Ahiram, Abi-baal, and Ethbaal (Ittobaʿal) in the 10th century, as excavations have shown. The history of this time period is mainly a history of Tyre, which not only rose to a hegemony among the Phoenician states but also founded colonies beyond the seas. Unfortunately, the native historical records of the Phoenicians have not survived, but it is clear from the Bible that the Phoenicians lived on friendly terms with the Israelites. In the 10th century Hiram, king of Tyre, built the Temple of Solomon at Jerusalem in return for rich gifts of oil, wine, and territory. In the following century Ethbaal of Tyre married his daughter Jezebel to Ahab, king of Israel, and Jezebel's daughter in turn married the King of Judah.

The Assyrian intrusion

In the 9th century, however, the independence of Phoenicia was increasingly threatened by the advance of Assyria. In 868 BC Ashurnasirpal II reached the Mediterranean and exacted tribute from the Phoenician cities. His son, Shalmaneser III, took tribute from the Tyrians and Sidonians and established a supremacy over Phoenicia, at any rate in theory, which was acknowledged by occasional payments of tribute to him and his successors. In 734 BC Tiglath-pileser III in his western campaign established his authority over Byblos, Arados, and Tyre. A fresh invasion, by Shalmaneser V, took place in 725 when he was on his way to Samaria, and in 701 Sennacherib, facing a rebellion of Philistia, Judah, and Phoenicia, drove out and deposed Luli, identified as king of both Sidon and Tyre. In 678 Sidon rebelled against the Assyrians, who marched down



Ruins of the Temple of Bacchus at Baalbek.
Peter Fenwick—The J. Allan Cash Photolibrary

and annihilated the city, rebuilding it on the mainland. Sieges of Tyre took place in 672 and 668, but it resisted both, only submitting in the later years of Ashurbanipal.

During the period of Neo-Babylonian power, which followed the fall of Nineveh in 612 BC, the pharaohs made attempts to seize the Phoenician and Palestinian seaboard. Nebuchadrezzar II, king of Babylon, having sacked Jerusalem, marched against Phoenicia and besieged Tyre, but it held out successfully for 13 years, after which it capitulated, seemingly on favourable terms.

Persian period. Phoenicia passed from the suzerainty of the Babylonians to that of their conquerors, the Persians, in 538 BC. Not surprisingly, the Phoenicians turned as loyal supporters to the Persians, who had overthrown their oppressors, and reopened to them the trade of the Orient. Lebanon, Syria–Palestine, and Cyprus were organized as the fifth satrapy (province) of the Persian Empire. At the time of Xerxes I's invasion of Greece, the city of Sidon was considered to be the principal city of Phoenicia, and the ships of Sidon were considered the finest part of Xerxes' fleet, its king ranking next to Xerxes and before the king of Tyre. (Phoenician coins have been used to supplement historical sources on the period. From the reign of Darius I [522–486 BC], the Persian monarchs had allowed their satraps and vassal states to coin silver and copper money. Arados, Byblos, Sidon, and Tyre therefore issued a coinage of their own.) In the 4th century Tyre, and later Sidon, revolted against the Persian king. The revolt was suppressed in 345 BC.

Greek and Roman periods. In 332 BC Tyre resisted Alexander the Great in a siege of eight months. Alexander finally captured the city by driving a mole into the sea from the mainland to the island. As a result Tyre, the inhabitants of which were largely sold into slavery, lost all importance, soon being replaced in the leadership of the Oriental markets by Alexandria, the conqueror's newly founded city in Egypt. In the Hellenistic period (323–30 BC) the cities of Phoenicia became the prize for the competing Macedonian dynasties, controlled first by the Ptolemies of Egypt in the 3rd century BC and then by the Seleucids of Syria in the 2nd and early decades of the 1st century BC. The Seleucids apparently permitted a good measure of autonomy to the Phoenician cities. Tigranes II of Armenia brought an end to the Seleucid dynasty in 83 BC and extended his realm to Mt. Lebanon. The Romans eventually intervened to restore Seleucid sovereignty, but when anarchy prevailed they imposed peace and assumed direct rule in 64 BC.

Phoenicia was incorporated into the Roman province of Syria, though Aradus, Sidon, and Tyre retained self-government. Berytus (Beirut), relatively obscure to this point, rose to prominence by virtue of Augustus' grant of Roman colonial status and by the lavish building program financed by Herod the Great (and in turn by his grandson and great-grandson). Under the Severan dynasty (AD 193–235) Sidon, Tyre, and probably Heliopolis (Baalbek) also received colonial status. Under this dynasty the province of Syria was partitioned into two parts: Syria Coele ("Hollow Syria"), comprising a large region loosely defined as north and east Syria; and Syria Phoenice in the southwestern region, which included not only coastal Phoenicia but also the territory beyond the mountains and into the Syrian desert. Under the provincial reorganization of the Eastern Roman emperor Theodosius II in the early 5th century AD, Syria Phoenice was expanded into two provinces: Phoenice Prima (Maritima), basically ancient Phoenicia; and Phoenice Secunda (Libanesia), an area extending to Mt. Lebanon on the west and deep into the Syrian desert on the east. Phoenice Secunda included the cities of Emesa (its capital), Heliopolis, Damascus, and Palmyra.

During the period of the Roman Empire the native Phoenician language died out in Lebanon and was replaced by Aramaic as the vernacular. Latin, the language of the soldiers and administrators, in turn fell before Greek, the language of letters of the eastern Mediterranean, by the 5th century AD. Lebanon produced a number of important writers in Greek, most notably Philo of Byblos (64–141), and in the 3rd century Porphyry of Tyre and Iamblichus of Chalcis in Syria Coele. Porphyry played a key role in disseminating the Neoplatonic philosophy of his master Plotinus, which would influence both pagan and Christian thought in the later Roman Empire.

In many respects, the two most important cities of Lebanon during the time of the Roman Empire were Heliopolis and Berytus. At Heliopolis the Roman emperors, particularly the Severans, constructed a monumental temple complex, the most spectacular elements of which were the Temple of Jupiter Heliopolitanus and the Temple of Bacchus. Berytus, on the other hand, became the seat of the most famous provincial school of Roman law. The school, which probably was founded by Septimius Severus, lasted until the destruction of Berytus itself by a sequence of earthquakes, tidal wave, and fire in the mid-6th century. Two of Rome's most famous jurists, Papinian and Ulpian, both natives of Lebanon, taught as professors at the law school under the Severans. Their judicial opinions constitute well over a third of the Pandects (Digest) contained in the great compilation of Roman law commissioned by the emperor Justinian I in the 6th century AD.

In 608–609 the Persian king Khosrow II pillaged Syria and Lebanon and reorganized the area into a new satrapy, excluding only Phoenicia Maritima. Between 622 and 629 the Byzantine emperor Heraclius mounted an offensive and restored Syria–Lebanon to his empire. This success was short-lived; in the 630s Muslim Arabs conquered Palestine and Lebanon, and the old Phoenician cities offered only token resistance to the invader.

(R.D.B./W.L.O./G.R.B.)

LEBANON IN THE MIDDLE AGES

The population of Lebanon did not begin to take its present form until the 7th century AD. At some time in the Byzantine period, a military group of uncertain origin, the Mardaites, established themselves in the north among the indigenous population. From the 7th century onward another group entered the country, the Maronites, a Christian community adhering to the Monothelite doctrine. Forced by persecution to leave their homes in northern Syria, they settled in the northern part of the mountain and absorbed the Mardaites and indigenous peasants to form the present Maronite Church. Originally Syriac-speaking, they gradually adopted the Arabic language although keeping Syriac for liturgical purposes. In south Lebanon Arab tribesmen came in after the Muslim conquest of Syria in the 7th century and settled among the indigenous people. In the 11th century many were converted to the Druze faith, an esoteric offshoot of Shī'ite Islām. South Lebanon became the headquarters of the faith. Groups of Shī'ite Muslims settled on the northern and southern fringes of the mountain and in al-Biqā'. In the coastal towns the population became mainly Sunnite Muslim, but in town and country alike there remained considerable numbers of Christians of various sects. In course of time, virtually all sections of the population adopted Arabic, the language of the Muslim states in which Lebanon was included.

The Maronites

Beirut and Mt. Lebanon were ruled by the Umayyads (661–750) as part of the district of Damascus. Despite the occasional rising by the Maronites, Lebanon provided naval forces to the Umayyads in their interminable warfare with the Byzantines. The 8th-century Beirut legist al-Awzā'ī established a school of Islāmic law that heavily influenced Lebanon and Syria. From the 9th to the 11th century coastal Lebanon was usually under the sway of independent Egyptian Muslim dynasties, although the Byzantine Empire attempted to gain portions of the north.

At the end of the 11th century Lebanon became a part of the crusaders' states, the north being incorporated in the county of Tripolis, the south in the kingdom of Jerusalem. The Maronite Church began to accept papal supremacy, while keeping its own patriarch and liturgy.

Despite the strong fortresses of the crusaders, a Muslim reconquest of Lebanon, under the leadership of Egypt, began with the fall of Beirut in 1187. Mongol raids against al-Biqā' valley were defeated. Lebanon became part of the Mamlūk state of Egypt and Syria in the 1280s and 1290s and was divided among several provinces. Mamlūk rule, which allowed limited local autonomy to regional leaders, encouraged commerce. The coastal cities, and especially Tripoli, flourished, and the people of the interior were left largely free to manage their own affairs.

OTTOMAN PERIOD

Ottoman expansion began in the area under Selim I (1512–20). He defeated the Mamlūks in 1516–17 and added Lebanon (as part of Mamlūk Syria and Egypt) to his empire. Between the 16th and 18th centuries Ottoman Lebanon evolved a social and political system of its own. Ottoman Aleppo or Tripoli governed the north, Damascus the centre, and Sidon (after 1660) the south. Coastal Lebanon and al-Biqā' valley were usually ruled more directly by Istanbul, while Mt. Lebanon enjoyed semiautonomous status. The population took up its present position: the Shī'ites were driven out of the north but increased their strength in the south; many Druze moved from south Lebanon to Jebel Druze (Jabal ad-Durūz) in southern Syria; Maronite peasants, increasing in numbers, moved south into districts mainly populated by Druze. Monasteries acquired more land and wealth. In all parts of the mountains there grew up families of notables, who controlled the land and established a feudal relation with the cultivators; some were Christian, some Druze, who were politically dominant. From them arose the House of M'an, which established a princedom over the whole of Mt. Lebanon and was accepted by Christians and Druze alike. Fakhr ad-Dīn II ruled most of Lebanon from 1593 to 1633 and encouraged commerce. When the House of Ma'n died out in 1697, the notables elected as prince a member of the Shihāb family who were Sunnite Muslims but with Druze followers, and this family ruled until 1842. Throughout this period European influence was growing. European trading colonies were established in Saïda and other coastal towns, mainly to trade in silk, the major Lebanese export from the 17th to the 20th century. French political influence was great, particularly among the Maronites, who formally united with the Roman Catholic Church in 1736.

European influence

The 19th century was marked by economic growth, social change, and political crisis. The growing Christian population moved southward and into the towns, and toward the end of the century many of these Christians emigrated to North America, South America, and Egypt. French Catholic and U.S. Protestant mission schools, as well as schools of the local communities, multiplied: in 1866 the American mission established the Syrian Protestant College (later the American University of Beirut), and in 1881 the Jesuits started the Université Saint-Joseph. Such schools produced a literate class, particularly among the Christians, that found employment as professionals. Beirut became a great international port, and its merchant houses established connections with Egypt, the Mediterranean countries, and England.

The growth of the Christian communities upset the traditional balance of Lebanon. The Shihāb princes inclined more and more toward them, and part of the family indeed became Maronites. The greatest of them, Bashīr II (reigned 1788–1840), after establishing his power with the help of Druze notables, tried to weaken them. When the Egyptian troops of Ibrāhīm Pasha occupied Lebanon and Syria in 1831, Bashīr formed an alliance with him to limit the power of the ruling families and to preserve his own power. But Egyptian rule was ended by Anglo-Ottoman intervention, aided by a popular rising in 1840, and Bashīr was deposed. With him the princedom virtually ended; his weak successor was deposed by the Ottomans in 1842, and from that time relations grew worse between the Maronites, led by their patriarch, and the Druze, trying to retain their traditional supremacy. The French supported the Maronites and the British supported a section of the Druze, while the Ottoman government encouraged the collapse of the traditional structure, which would enable it to impose its own direct authority. The conflict culminated in the massacre of Maronites by the Druze in 1860. The complacent attitude of the Ottoman authorities led to direct French intervention on behalf of the Christians. The powers jointly imposed the Organic Regulation of 1861 (modified in 1864), which gave Mt. Lebanon, the axial mountain region, autonomy under a Christian governor appointed by the Ottoman sultan, assisted by a council representing the various communities. Mt. Lebanon prospered under this regime until World War I, when the Ottoman government placed it under strict control, similar to that already established for the coast and al-Biqāʿ valley.

French intervention

FRENCH MANDATE

At the end of the war Lebanon was occupied by Allied forces and placed under a French military administration. In 1920 Beirut and other coastal towns, al-Biqāʿ, and certain other districts were added to the autonomous territory Mt. Lebanon as defined in 1861, to form Greater Lebanon (Grand Liban; subsequently called the Lebanese Republic). In 1923 the League of Nations formally gave the mandate for Lebanon and Syria to France. The Maronites, strongly pro-French by tradition, welcomed this, and during the next 20 years, while France held the mandate, the Maronites were favoured. The expansion of prewar Lebanon into Greater Lebanon, however, changed the balance of the population. Although the Maronites were the largest single element, they no longer formed a majority. The population was more or less equally divided between Christians and Muslims, and a large section of it wanted neither to be ruled by France nor to be part of an independent Lebanon, but rather to form part of a larger Syrian or Arab state. To ease tensions between the communities, the constitution of 1926 provided that each should be equitably represented in public offices. Thus by convention the president of the republic was normally a Maronite, the prime minister a Sunnite Muslim, and the speaker of the chamber a Shīʿite Muslim.

French administration was reasonably efficient. Public utilities and communications were improved, and education was expanded (although higher education was left almost wholly in the hands of religious bodies). Beirut prospered as a centre of trade with surrounding countries, but agriculture was depressed by the decline of the silk industry and the worldwide economic depression. As the middle class of Beirut grew and a real if fragile sense of common national interest sprang up alongside communal loyalties, there grew also the desire for more independence. A Franco-Lebanese treaty of independence and friendship was signed in 1936 but was not ratified by the French government. Lebanon was controlled by the Vichy authorities after the fall of France in 1940 but was occupied by British and Free French troops in 1941. The Free French representative proclaimed the independence of Lebanon and Syria, which was underwritten by the British government. Because of their own precarious position, however, the Free French were unwilling to relax control. In 1943, however, they held elections, which resulted in victory for the Nationalists. Their leader, Bishara al-Khuri, was elected president. The new government passed legislation introducing certain constitutional changes that eliminated all traces of French influence, to which the French objected. On Nov. 11, 1943, the President and almost the entire government were arrested by the French. This led to an insurrection followed by British diplomatic intervention; the French restored the government and transferred powers to it. After another crisis in 1945, agreement was reached on a simultaneous withdrawal of British and French troops. This was completed by the end of 1946, and Lebanon became wholly independent; it had already become a member of the United Nations and the Arab League. (R.D.B./W.L.O.)

LEBANON AFTER INDEPENDENCE

For many years Lebanon maintained its parliamentary democracy, despite serious trials. The main problem for Lebanon was to implement the unwritten power-sharing National Pact of 1943 between the Christians and Muslims. In the early years of independence, so long as no urgent call for pan-Arab unity came from outside, the National Pact faced no serious strains.

The Khuri regime, 1943–52. The Maronite president Bishara al-Khuri closely cooperated with the Sunnite leader Riad as-Sulh, who was premier most of the time. A temporary amendment of the constitution permitted the president, in 1949, a second six-year term. The parliamentary elections of 1947 were blatantly rigged to produce a Parliament favourable to the amendment. This, together with the open favouritism of the President toward his friends and the gross corruption he allegedly condoned, made Khuri increasingly unpopular after his reelection in 1949.

The military coup that overthrew the Kuwatli regime in Syria in March 1949 encouraged the opponents of Khuri in Lebanon. In July 1949 the Syrian Social Nationalist Party (PPS) tried to overthrow the regime by force. The coup failed and its leaders were seized and shot. The PPS took its revenge by securing the assassination of Khuri's premier in 1951. The mounting opposition to the Khuri regime culminated in September 1952 in a general strike that forced his resignation. Camille Chamoun was elected by the Parliament to succeed him.

The Chamoun regime and the 1958 crisis. The presidency of Camille Chamoun coincided with the rise of Nasser in Egypt. During the Suez war (October–December 1956), Chamoun earned Nasser's enmity by refusing to break off diplomatic relations with Britain and France, which had joined Israel in attacking Egypt. Chamoun was accused of seeking to align Lebanon with the Western-sponsored Baghdad Pact.

Lebanon's difficult position

Matters came to a head following the parliamentary elections of 1957, which allegedly were manipulated to produce a Parliament favourable to the reelection of Chamoun. When Syria entered into a union with Egypt, as the United Arab Republic, in February 1958, the Muslim opposition to Chamoun in Lebanon hailed the union as a triumph for pan-Arabism, and there were widespread demands that Lebanon be associated in the union. In May a general strike was proclaimed, and the Muslims of Tripoli rose in armed insurrection. The insurrection spread, and the army was asked to take action against the insurgents. The commanding general, Fuad Chehab, refused to attack them for fear that the army, which was composed of Christians and Muslims, would split apart. The Chamoun government took the issue of external intervention to the United Nations, accusing the United Arab

Republic of intervention, and UN observers were sent to Lebanon. When in July the pro-Western regime in Iraq was toppled in a coup, President Chamoun immediately requested U.S. military intervention, and on the following day U.S. marines landed outside Beirut. The presence of U.S. troops had little immediate effect on the internal situation, but the insurrection slowly faded out. Parliament turned to the commander of the army, General Chehab, to succeed Chamoun as his term ended; Rashid Karami became the new premier.

The Chehab, Hélou, and Franjieh regimes, 1958–76. The crisis had been resolved by compromise, and the Chehab regime was successful in maintaining the compromise and promoting the national unity of the Lebanese people. By his refusal as army commander to take offensive action against the insurgents in 1958, Chehab had earned the confidence of the Muslims. Once in power, he proceeded to allay long-standing Muslim grievances by associating Muslims more closely in the administration and by attending to neglected areas of Lebanon where Muslims predominated. Internal stability was further promoted by the maintenance of good relations with the United Arab Republic, which, even after the Syrian secession in 1961, remained highly popular with the Muslim Lebanese. The economic boom that had begun under the Chamoun regime as the result of the flight of capital from the unstable Arab world into Lebanon continued under the Chehab regime.

National unity of the Lebanese people

Charles Hélou, a former journalist and member of Khuri's Constitutional Bloc, was elected to succeed Chehab in 1964. His government, essentially a weaker version of the Chehab regime, was followed in 1970 by the troubled regime of Suleiman Franjieh.

Despite the apparent calm of the years 1958–69, events in Lebanon moved toward the outbreak of one of the most destructive civil wars in modern history. The essential issues were separate but strongly and immediately affected one another. First came the fact that the political structure of Lebanon was based on a sort of floating consensus—an agreement among leaders of the various religious and ethnic minorities over their respective roles in the state. Under the regimes of Chehab and Hélou, the personal and group ambitions of the various sectors of society were held in check, and the organs of the state, especially the army and the security police, were employed to prevent recourse to violence.

During that period, major changes took place in Lebanese society, most significantly urbanization, which brought 40 percent of the Lebanese population to the city of Beirut. But the city, like the country, failed to make its people homogeneous; Beirut became a reflection of Lebanon as a whole. Each quarter took on a religious affiliation, with virtually every village having an enclave established in one suburb or more. And the newcomers suffered from deep and growing social and economic contrasts with their more affluent neighbours. Those who remained in the rural areas lost even more ground, relatively speaking. By the early 1970s, agriculture, in which nearly half of the population was employed, accounted for less than 11 percent of the country's gross domestic product, while the share of urban commerce was rising. An already existing schism was deepened between the groups, which increasingly took on an urban Christian versus rural Muslim coloration.

Lebanese social change

A second factor, the role of Lebanon in the Arab world, was also a complex issue. Many Syrians still felt that the French decision to separate Lebanon from Syria in 1920 was invalid; many in Lebanon agreed. Lebanon's noninvolvement in the Arab-Israeli Wars of 1967 and 1973, its strong and often heavy-handed security policies, and rumours of its secret understandings with Israel all directed attention on this issue.

Third, the Palestinians thought of Lebanon, after the ruinous Jordanian campaign against them in September 1970, as their last refuge, and by 1973 roughly one person in each 10 in Lebanon was a Palestinian. Yet the Palestinians were constantly made aware of their separate and inferior status; landless and mostly poor, they were exploited as a source of cheap labour. Increasingly their politics became radicalized, and they found common cause with those Lebanese who were poor, rural, and mainly Muslim. As they acquired structure, motivation, and arms during the period after the fall of their base in Jordan, they were sought out as allies by groups in Lebanon. Faced in September 1975 by an Egyptian-Israeli interim agreement on Sinai, the Palestinians concluded that they were being deserted by the Arab states and would shortly be suppressed in Lebanon.

The civil war, 1975–76. *Beginning of the war.* Toward the end of the presidency of Charles Hélou, the various factions of the Palestine Liberation Organization (PLO) began to clash with Lebanese security forces. Under an agreement announced in Cairo on Nov. 3, 1969, the Lebanese government gave the Palestinians virtually a free hand in the refugee camps and at forward posts in the south along the Israeli frontier. In return, the PLO promised not to intervene in Lebanese politics; this was difficult for the Palestinians and not desired by the Lebanese left. When the Lebanese failed to restrain the Palestinians, the Israelis began to raid the south with increasing severity, and this encouraged the Lebanese Christian right, particularly the militant Phalangist Party, to attack the Palestinians with its well-organized and well-armed militia.

It was in this atmosphere that Suleiman Franjieh was elected president on Aug. 17, 1970. Franjieh, however, was not able to solve the two basic political and foreign problems that troubled the country: should more power in the Lebanese government be given to the Shīʿites and other Muslims who had become a majority of the population, and should Lebanon support or suppress the PLO? Events moved rapidly toward civil war, and by 1975 the mostly Muslim Lebanese National Movement led by Kamal Jumblatt sought political reform and support for the Palestinian guerrillas. Into this arena stepped the relatively deprived Shīʿite Muslims, by now the most numerous religious community in Lebanon. Maronite Christians, intent on preserving their concept of Lebanon, frantically sought to keep their political dominance by crushing the power of the Lebanese leftists and particularly the PLO, whose actions seemed (from the perspective of many Maronites) to threaten the unity and safety of the nation.

Hardly a day passed after the beginning of full civil war in April 1975 without a battle somewhere in Lebanon. The country was torn apart, and the central government virtually ceased to exist. The army, long the mainstay of the government, largely dissolved while the combatants, amply supplied by various foreign groups, turned upon one another with a ferocity—and firepower—almost unequaled in such a small area of the world.

Final phases of the war. Gradually the left and the Palestinians began to win the war. By the early months of 1976, it seemed clear that the Christians were losing and that either they would be defeated (so that Lebanon would be reconstituted as a left-dominated, pro-PLO state) or Lebanon would be partitioned. Either case appeared to the Syrians likely to bring Israeli intervention. This realization forced a reversal of Syrian policy, ending in Pres. Ḥafiz al-Assad's support for the Christians. Ironically, both the Syrians and the Israelis, so opposed to one another on other issues, took up the cause of the Lebanese Christians. Syria prevented the Palestinians from taking strategic points, while Israel blockaded the coast, trained a contingent of Lebanese in Israel, and shipped equipment to the Christian sector. During the summer of 1976 Syrian military units entered the country from the east with about 20,000 soldiers. The Christians, with strong support from Syria, began to win the civil war as they attacked Palestinian refugee camps.

Syrian reversal of position

Parliament elected a new president, Elias Sarkis, on May 8, 1976, with the support of Syria, but he was not inaugurated until September 23. Meanwhile, Lebanon was effectively partitioned along the "Green Line," which passed through the centre of Beirut (east–west) and along the main road to Damascus; to the north was a Christian government, to the south a leftist (Druze-Muslim-Palestinian) government led by Kamal Jumblatt until he was murdered in March 1977. Repeated attempts were made to bring the fighting to an end, until finally a formal "summit" meeting, held on Oct. 25–26, 1976, established an Arab

League peacekeeping force of 30,000 troops, who were mostly Syrian. By the end of November, despite continued minor clashes, the intensity of the civil war decreased, but it did not really end, since the major issues that had caused the fighting to start in 1975 remained unresolved.

Continued fighting among the Lebanese factions led to the loss of prestige of the former political elite and to the emergence of a new generation of militia leaders, except in the Phalangist Party, whose existing leadership dominated the Christian-rightist coalition so successfully that the Syrian army of occupation once again began to support the Muslim-leftist-Palestinian groups in 1978. The destruction and violence had caused hundreds of thousands of Lebanese to flee their homes in southern Lebanon, where the threat of Israeli intervention stopped Syria from imposing a peace. Israel invaded the area on March 14–15, 1978, to destroy Palestinian bases and to force Lebanon to curb raids by the PLO into Israel. A small contingent of UN forces replaced the Israelis by June; Israel continued, however, to supply arms, money, and troops to the Christians in the south, while the Palestinians soon returned to the same region.

Consequences of the war. The civil war was a catastrophe for the Lebanese, whose country lay in ruins. There seemed to be no compromise acceptable to the Muslims, who numbered more than one-half of the population, and to the Christians, who were determined to keep their control of key government institutions. Foreign intervention merely restrained open, full-scale warfare. Economic destruction was massive, but this was overcome to a certain extent by increased remittances from Lebanese working abroad during the boom years in the oil-producing countries. From 1975 to 1982, while tourism and industrial production declined sharply, capital invested in real estate, banking, and the newly decentralized commercial and service sectors helped compensate for economic losses. The chief political problem was the bitterness caused by the thousands of deaths and the ensuing hatreds that promised to destroy the possibility of Lebanese living together again in one nation with one government.

Economic effects of the war

The war left the Palestinians with perhaps 20,000 killed and twice that many wounded. The Syrians appeared stronger than before, but, having got into Lebanon, they faced the problem of extricating themselves. Only Israel among the states of the Middle East appeared to have "won." The Palestinians lost their major bid, Syria feared Israeli intervention, and the Lebanese Christians were in Israel's debt. More important, the horror of the war had caused Arabs everywhere to question, as never before, the very dream of pan-Arabism.

The Israeli invasion of 1982. The political disintegration of Lebanon led directly to intensified external intervention. Bashir Gemayel, the leader of the Phalangist militia, whose strength derived in part from extensive Israeli aid, forcibly united under his control all the Maronite private armies and thereby created a ministate in East Beirut and the northern coastal sector of Lebanon. The Syrian army was dominant in most of the rest of Lebanon, but a jumble of factions, many of which were armed and paid by outsiders, disputed Syria's power and wreaked havoc because of their internecine quarrels.

Israeli forces bombed PLO headquarters in West Beirut on July 17, 1981, causing in the process more than 300 civilian deaths. This attack led the United States to arrange a cease-fire between the Israelis and the PLO, which, it was hoped, would end raids into northern Israel. The situation, however, erupted on June 6, 1982, when an estimated 60,000 Israeli troops invaded Lebanon.

Although the stated goal of Israel was only to secure the territory north of its border with Lebanon so as to stop PLO raids, Israeli Prime Minister Menachem Begin sought to destroy the PLO and establish in power a Lebanese government that would conclude a peace treaty with Israel along the lines of the Egyptian-Israeli peace of 1979. The invasion was successful, as Syrian forces were defeated, the PLO retreated to West Beirut, and Egypt and the other Arab states did little but protest. From late June to August, Israel hesitated to attack PLO and leftist Muslim troops in densely populated West Beirut. Instead, Israel shelled, bombed, and blockaded the area to pressure the PLO and Syrian garrisons to evacuate their forces.

Under supervision by an international (U.S., French, and Italian) force, PLO leaders and troops left Beirut for a number of Arab countries in late August. Because Syria supported the PLO forces remaining in northern Lebanon and in al-Biqāʿ valley, the forces could not be compelled by Israel to leave, but the Syrian backing was used to foster a PLO leadership that opposed the PLO chairman, Yāsir ʿArafāt. (In heavy fighting near Tripoli, ʿArafāt was forced into exile in December 1983 for a second time, on this occasion at the instigation of the Syrians.) The Israeli victory in the south and centre was shared by the Phalangists, who then had no barrier to electing their leader president of Lebanon. Bashir Gemayel, however, was assassinated before his inauguration. The Phalangists then secured the election of Bashir's brother, Amin Gemayel, to replace the exhausted and ineffectual Sarkis as president. After West Beirut was occupied by the Israelis, Phalangist militiamen massacred perhaps as many as 1,000 Palestinians in two refugee camps in Beirut in revenge for the death of Bashir Gemayel.

On May 17, 1983, Israel and Lebanon concluded what was very nearly a peace treaty. It called for the withdrawal of Israeli forces, a special security zone in the south, and the establishment of bilateral relations. Israel's power in Lebanon deteriorated as growing opposition from various Lebanese groups resulted in armed attacks on Israelis, and Israeli casualties mounted; in September 1983 Israel began withdrawing its forces. The international peacekeeping force left Beirut in February 1984 after suffering heavy casualties, and in March Syria and Lebanese Muslims and leftists forced Pres. Amin Gemayel to abrogate the Lebanon-Israel agreement. By June 1985 Israel had withdrawn its military from most of Lebanon. This abrupt reversal among the intervening foreign states exacerbated political instability inside Lebanon. The Christian and rightist movement, the Shīʿite-Druze alliance, and the PLO all split asunder over the question of accepting or rejecting Syria's leadership.

President Gemayel rejected in 1986 the Syrian-arranged compromise proposal backed by the leftist militias, whose power was weakened by the fighting between Amal (a Shīʿite political and military organization) and the PLO. The amazingly resilient Lebanese economy itself finally began to collapse under the cumulative strain of years of warfare and destruction, as the value of Lebanese currency drastically declined and public services in the country deteriorated. When Gemayel's term ended on Sept. 22, 1988, Parliament could not agree on the selection of a new president; instead, Gemayel named Gen. Michel Aoun as prime minister, despite the continuing claim to that office by the incumbent, Salim al-Hoss. Lebanon thus had no president but two prime ministers, and the complete partition of the country seemed inevitable.

Lebanese of nearly all factions and groups rejected the possible disappearance of their country. Instead, the chief issue became which one of the groups would dominate a newly reunited Lebanon. In March 1989 General Aoun launched a "war of liberation" against Syria and its Lebanese allies; despite Iraq's covert assistance, this war failed, and in September Aoun accepted a cease-fire.

On October 22 most members of the Lebanese Parliament (last elected in 1972) met in Ṭāʾif, Saudi Arabia, and accepted a constitutional compromise that adjusted the Parliament, presidency, and cabinet so that Christian and Muslim representatives would equally share power. On November 5 the old Parliament elected René Moawad as president. Moawad, however, was assassinated on November 22, and, though Elias Hrawi was elected two days later, General Aoun denounced both presidential elections as invalid since the whole process of political compromise ignored the issue of Syrian withdrawal from Lebanon. After more factional fighting in early 1990, Syria finally took decisive action against Aoun—on October 13 the Lebanese government's central army and the Syrians forced his surrender. President Hrawi then embarked upon the delicate and dangerous process of consolidating and extending the power of the Lebanese government, first by disarming the

Surrender of General Aoun

militias in Beirut and then by reaching out into other parts of the country. (W.L.O.)

From the beginning of the civil war in 1975 to the early 1990s, perhaps as many as 150,000 Lebanese died in the various types of fighting. About one-quarter of the country's population fled abroad, and hundreds of thousands were forced to move from one part of Lebanon to another. The Lebanese were exhausted by the interminable violence, and most seemed prepared to accept the compromise peace that continued in Lebanon throughout the 1990s, as Sunnite, Shī'ite, and Christian factions vied for political power within Lebanon's revived constitutional framework.

Although small pockets of violence continued in the country throughout that decade (particularly along the Israeli border), the presence of a large number of Syrian troops within the country, although unwelcome to many Lebanese, served to bolster Lebanon's central government. Reconstruction of the country, particularly war-torn Beirut, began apace during the 1990s, as the Lebanese government, particularly its energetic prime minister, Rafīq al-Ḥarīrī, sought substantial investment from abroad to revitalize the country's shattered economy, once the healthiest in the region.

Noteworthy also in the final decade of the 20th century was the rise to prominence of the Shī'ite group Hezbollah. Founded during the early 1980s, the pro-Iranian group aspired to eliminate Israeli influence from the country and replaced the PLO as Israel's principal antagonist in southern Lebanon, waging a vigorous war against the Jewish state even after that country's final withdrawal from Lebanon in mid-2000. Seeking to develop a broader base in the 1990s, Hezbollah became increasingly active in Lebanon's coalition politics and established its own social, medical, and educational infrastructure to serve its supporters. (Ed.)

For later developments in the political history of Lebanon, see the BRITANNICA BOOK OF THE YEAR.

BIBLIOGRAPHY. General discussions of the land and people may be found in W.B. FISHER, *The Middle East,* 7th ed. (1978); and DAVID C. GORDON, *The Republic of Lebanon: Nation in Jeopardy* (1983). SHEREEN KHAIRALLAH, *Lebanon* (1979), is an annotated bibliography of works on all aspects of the country. Economic and social matters are discussed by ABDUL-AMIR BADRUD-DIN, *The Bank of Lebanon* (1984); NADIM G. KHALAF, *The Economic Implications of the Size of Nations; with Special Reference to Lebanon* (1971); HUDA C. ZURAYK and HAROUTUNE K. ARMENIAN, *Beirut 1984* (1985); JOSEPH CHAMIE, *Religion and Fertility: Arab Christian-Muslim Differentials* (1981); YUSIF A. SAYIGH, *Entrepreneurs of Lebanon* (1962), a study of the role of entrepreneurs in the national development of Lebanon; ANNE H. FULLER, *Buarij: Portrait of a Lebanese Muslim Village* (1961, reissued 1968); LILIANE GERMANOS-GHAZALY, *Le Paysan, la terre et la femme: organisation sociale d'un village du Mont-Liban* (1978); FRIEDRICH RAGETTE (ed.), *Beirut of Tomorrow: Planning for Reconstruction* (1983); and JOHN GULICK, *Tripoli: A Modern Arab City* (1967). Useful discussions of Lebanese government include MICHAEL W. SULEIMAN, *Political Parties in Lebanon: The Challenge of a Fragmented Political Culture* (1967); and GEORGE GRASSMUCK and KAMAL SALIBI, *Reformed Administration in Lebanon,* 2nd ed. (1964). See also ADEL A. FREIHA, *L'Armée et l'état au Liban, 1945–1980* (1980); R.D. MCLAURIN, "Lebanon and Its Army: Past, Present, and Future," in EDWARD E. AZAR *et al., The Emergence of a New Lebanon: Fantasy or Reality?,* pp. 79–114 (1984); WILLIAM W. HARRIS, *Faces of Lebanon: Sects, Wars, and Global Extensions* (1997); and CHARLES WINSLOW, *Lebanon: War and Politics in a Fragmented Society* (1996). Cultural matters are discussed by LAWRENCE I. CONRAD, "Culture and Learning in Beirut," *The American Scholar,* 52:463–478 (Autumn 1983); and FRIEDRICH RAGETTE, *Architecture in Lebanon: The Lebanese House During the 18th and 19th Centuries* (1974, reprinted 1980).

Ancient history is detailed in *The Cambridge Ancient History,* especially vol. 1 in 2 pts., 3rd ed. (1970–71), vol. 2, part 1, 3rd ed. (1973), and vol. 3, part 3, 2nd ed. (1982); and in DONALD HARDEN, *The Phoenicians,* rev. ed. (1971). Other useful studies include MAURICE DUNAND, *Byblos: Its History, Ruins and Legends,* 2nd ed. (1968; originally published in French, 2nd ed., 1968, reissued 1973); FRIEDRICH RAGETTE, *Baalbek* (1980); F.M. HEICHELHEIM, "Roman Syria," in TENNEY FRANK (ed.), *An Economic Survey of Ancient Rome,* vol. 4, pp. 121–257 (1938, reprinted 1975); J.-P. REY-COQUAIS, "Syrie Romaine, de Pompée à Dioclétien," *Journal of Roman Studies,* 68:44–73 (1978); HILDEGARD TEMPORINI and WOLFGANG HAASE (eds.), *Aufstieg und Niedergang der Römischen Welt,* vol. 2, part 8, *Politische Geschichte: Provinzen und Rundvölker: Syrien, Palästina, Arabien,* pp. 3–294 (1977); and NINA JIDEJIAN, *Byblos Through the Ages* (1968), *Tyre Through the Ages* (1969), *Sidon Through the Ages* (1971), *Beirut Through the Ages* (1973), and *Baalbek: Heliopolis, City of the Sun* (1975).

The most important works on Lebanon's medieval and modern history are PHILIP K. HITTI, *Lebanon in History: From the Earliest Times to the Present,* 3rd ed. (1967); and KAMAL S. SALIBI, *The Modern History of Lebanon* (1965, reissued 1977), and *A House of Many Mansions: The History of Lebanon Reconsidered* (1988). The Ottoman period is discussed by ABDUL-RAHIM ABU-HUSAYN, *Provincial Leaderships in Syria, 1575–1650* (1985); DOMINIQUE CHEVALLIER, *La Société du Mont Liban à l'époque de la révolution industrielle en Europe* (1971, reissued 1982); and ILIYA F. HARIK, *Politics and Change in a Traditional Society: Lebanon, 1711–1845* (1968).

Twentieth-century history is explored by ALBERT H. HOURANI, *Syria and Lebanon: A Political Essay* (1946, reprinted 1968); and MICHAEL C. HUDSON, *The Precarious Republic: Political Modernization in Lebanon* (1968, reissued 1985). The civil war and subsequent events are evaluated by KAMAL S. SALIBI, *Cross Roads to Civil War: Lebanon, 1958–1976* (1976, reissued as *Crossroads to Civil War,* 1988); WALID KHALIDI, "Lebanon: Yesterday and Tomorrow," *The Middle East Journal,* 43(3):375–387 (Summer 1989); HELENA COBBAN, *The Making of Modern Lebanon* (1985); DAVID GILMOUR, *Lebanon, the Fractured Country,* rev. and updated ed. (1987); N. KLIOT, "The Collapse of the Lebanese State," *Middle Eastern Studies,* 23:54–74 (January 1987); HALIM BARAKAT (ed.), *Toward a Viable Lebanon* (1988); ROSEMARY HOLLIS and NADIM SHEHADI (eds.), *Lebanon on Hold: Implications for Middle East Peace* (1996); and HABIB C. MALIK, *Between Damascus and Jerusalem: Lebanon and Middle East Peace,* updated ed. (2000).

(S.G.K./C.F.M./R.D.B./W.L.O./G.R.B.)

The Evolution of Modern Western Legal Systems

The two great law families of modern Western civilization are civil law (also called Romano-Germanic law) and common law (also called Anglo-American law). They are descended from ancient Roman law and ancient Germanic tribal law and have been altered by various customary, ecclesiastical, feudal, commercial, and sociopolitical influences. A third law family, Socialist law, was in use primarily in the now-defunct Socialist states of eastern Europe.

Most of the laws of the legal systems of continental Europe are traditionally classified as civil law. This type of law spread to Latin America and, later, to those countries of Asia and Africa that found it necessary to westernize their laws, such as Japan (in which American law has also had great influence), Thailand, Turkey, and Ethiopia. It also prevails, supplemented by religious or customary laws, in those regions that were colonies, protectorates, or trust territories of France, Belgium, The Netherlands, Spain, Portugal, and Italy, including Morocco, Algeria, Zaire, Indonesia, and Somalia, as well as in some dependencies or outlying parts of European countries, such as Martinique and Curaçao in the West Indies. Civil law, supplemented increasingly by Islāmic law, has come to prevail in several countries of the Middle East and North Africa.

Common law is, essentially, the law of England and the law of those countries in which the law of England has been received or implanted. Although common law was often transformed in certain respects or supplemented by local or religious traditions, its principal features have been preserved. Common-law countries that have remained closest to the English tradition include Canada, Australia, New Zealand, the Republic of Ireland, and the present and former British areas of the West Indies. Common law also prevails in India, Pakistan, Myanmar (Burma), Malaysia, and Singapore, where it is supplemented in matters of personal status by religious laws. In Liberia and in most of those parts of Africa and Oceania that were British colonies, protectorates, or trust territories, common law is supplemented by native customs. The United States has developed a type of common law that differs in many respects from the English.

Civil law and common law have merged in the legal systems of several countries (Scotland, Puerto Rico, the Philippines, Israel, South Africa, Zimbabwe, Sri Lanka, Mauritius, and the Seychelles) and in other localized regions (Quebec, in Canada, and Louisiana, in the United States). In most of these places private law still follows a civil-law pattern. The laws of the Nordic countries (Sweden, Finland, Denmark, Norway, and Iceland) are also closer to civil law than to common law.

Socialist law—as practiced in present and former Socialist states—diverged from its Romano-Germanic roots in important respects, foremost in its collectivist spirit and its aim of assisting in the creation of a new social order. Nonetheless, many concepts of civil law are still used in Socialist legal thinking.

For coverage of related topics in the *Macropædia* and *Micropædia*, see the *Propædia*, Part Five, Division V, especially Section 551, and the *Index*.

This article is divided into the following sections:

The Western legal heritage

ROMAN LAW

In its strictest sense the term Roman law denotes the law of the city of Rome and of the Roman Empire that was in force during the period from the foundation of the city (traditional date 753 BC) until the fall of the Western Empire in the 5th century AD and the fall of the Eastern Empire in 1453. The term Roman law today, however, often refers to more than the laws of Roman society. The legal institutions evolved by the Romans had influence on the laws of other peoples in times long after the disappearance of the Roman Empire and in countries that were never subject to Roman rule. To take the most striking example, in a large part of Germany, until the adoption of a common code for the whole empire in 1900, the Roman law was in force as "subsidiary law"; that is, it was applied unless excluded by contrary local provisions. This law, however, which was in force in parts of Europe long after the fall of the Roman Empire, was not the Roman law in its original form. Although its basis was indeed the Corpus Juris Civilis—the codifying legislation of the emperor Justinian I—this legislation had been interpreted, developed, and adapted to later conditions by generations of jurists from the 11th century onward and had received additions from non-Roman sources.

The range of influence of Roman law

Development of the jus civile and jus gentium. In the great span of time during which the Roman Republic and Empire existed, there were many phases of legalistic development. During the period of the republic (753–31 BC), the *jus civile* (civil law) developed. Based on custom or legislation, it applied exclusively to Roman citizens. By the middle of the 3rd century BC, however, another type of law, *jus gentium* (international law), was developed by the Romans to be applied both to themselves and to foreigners. *Jus gentium* was not the result of legislation, but was, instead, a development of the magistrates and governors who were responsible for administering justice in cases in which foreigners were involved. The *jus gentium* became, to a large extent, part of the massive body of law that was applied by magistrates to citizens, as well as to foreigners, as a flexible alternative to *jus civile.*

Roman law, like other ancient systems, originally adopted the principle of personality—that is, that the law of the state applied only to its citizens. Foreigners had no rights and, unless protected by some treaty between their state and Rome, they could be seized like ownerless pieces of property by any Roman. But from early times there were treaties with foreign states guaranteeing mutual protection. Even in cases in which there was no treaty, the increasing commercial interests of Rome forced it to protect, by some form of justice, the foreigners who came within its borders. A magistrate could not simply apply Roman law because that was the privilege of citizens; even had there not been this difficulty, foreigners would probably have objected to the cumbersome formalism that characterized the early *jus civile.*

The law that the magistrates applied probably consisted of three elements: (1) an existing mercantile law that was used by the Mediterranean traders; (2) those institutions of the Roman law that, after being purged of their formalistic elements, could be applied universally to any litigant, Roman or foreigner; and (3) in the last resort, a magistrate's own sense of what was fair and just. This system of *jus gentium* was also adopted when Rome began to acquire provinces so that provincial governors could administer justice to the *peregrini* (foreigners). This word came to mean not so much persons living under another government (of which, with the expansion of Roman power, there came to be fewer and fewer) as Roman subjects who were not citizens. In general, disputes between members of the same subject state were settled by that state's own courts according to its own law, whereas disputes between provincials of different states or between provincials and Romans were resolved by the governor's court applying *jus gentium.* By the 3rd century AD, when citizenship was extended throughout the empire, the practical differences between *jus civile* and *jus gentium* ceased to exist. Even before this, when a Roman lawyer said that a contract of sale was *juris gentium,* he meant that it was formed in the same way and had the same legal results whether the parties to it were citizens or not. This became the practical meaning of *jus gentium.* Because of the universality of its application, however, the idea was also linked with the theoretical notion that it was the law common to all peoples and was dictated by nature—an idea that the Romans took from Greek philosophy.

Written and unwritten law. The Romans divided their law into *jus scriptum* (written law) and *jus non scriptum* (unwritten law). By "unwritten law" they meant custom; by "written law" they meant not only the laws derived from legislation but, literally, laws based on any written source.

There were various types of written law, the first of which consisted of *leges* (singular *lex*), or enactments of one of the assemblies of the whole Roman people. Although the wealthier classes, or patricians, dominated these assemblies, the common people, or plebeians, had their own council in which they enacted resolutions called *plebiscita.* Only after the passage of the Lex Hortensia in 287 BC, however, did *plebiscita* become binding on all classes of citizens; thereafter, *plebiscita* were generally termed *leges* along with other enactments. In general, legislation was a source of law only during the republic. When Augustus established the empire in 31 BC, the assemblies did not at once cease to function, but their assent to any proposal became merely a formal ratification of the emperor's wishes. The last known *lex* was passed during the reign of Nerva (AD 96–98).

Types of written law

The earliest and most important legislation, or body of *leges,* was the Twelve Tables, enacted in 451–450 BC during the struggle of the plebeians for political equality. It represented an effort to obtain a written and public code that patrician magistrates could not alter at will against plebeian litigants. Little is known of the actual content of the Twelve Tables; the text of the code has not survived, and only a few fragments are extant, collected from allusions and quotations in the works of authors such as Cicero. From the fragments it is apparent that numerous matters were treated, among them family law, delict (tort, or offense against the law), and legal procedure.

A second type of written law consisted of the *edicta* (edicts), or proclamations issued by a superior magistrate (praetor) on judicial matters. The office of praetor was created in 367 BC to take over the expanding legal work involving citizens; later, a separate praetor was created to deal with foreigners. Upon taking office, a praetor issued an edict that was, in effect, the program for his year in office. The curule aediles, who were the magistrates responsible for the care and supervision of the markets, also issued edicts. During the later stages of the republic, these praetorian and magisterial edicts became an instrument of legal reform, and *leges* ceased to be a major source of private law.

The Roman system of procedure gave the magistrate great powers for providing or refusing judicial remedies, as well as for determining the form that such remedies should take. The result of this magisterial system was the development of a new body of rules that existed alongside, and often superseded, the civil law. The *edicta* remained a source of law until about AD 131, when the emperor Hadrian commissioned their reorganization and consolidation and declared the resulting set of laws to be unalterable, except by the emperor himself.

Law-making power of magistrates

A third type of written law was the *senatus consulta,* or resolutions of the Roman senate. Although these suggestions to various magistrates had no legislative force during the republic, they could be given force by the magistrates' edicts. In the early empire, as the power of the assemblies declined and the position of the emperor increased, *senatus consulta* became resolutions that endorsed the proposals of the emperor. As the approval of the Senate became increasingly automatic, the emperor's proposals became the true instrument of power. Consequently, emperors ceased referring proposals to the Senate and, not long after the early imperial period, ended the practice of legislating through the Senate.

A fourth type of written law consisted of the *constitu-*

tiones principum, which were, in effect, expressions of the legislative power of the emperor. By the middle of the 2nd century AD, the emperor was, essentially, the sole creator of the law. The chief forms of imperial legislation were edicts or proclamations; instructions to subordinates, especially provincial governors; written answers to officials or others who consulted the emperor; and decisions of the emperor sitting as a judge.

The last type of written law was the *responsa prudentium,* or answers to legal questions given by learned lawyers to those who consulted them. Although law, written and unwritten, was originally a rather secretive monopoly of the college of pontiffs, or priests, a recognizable class of legal advisers, *juris consulti* or *prudentes,* had developed by the early 3rd century BC. These legal advisers were not professionals as such but men of rank who sought popularity and advancement in their public careers by giving free legal advice. They interpreted statutes and points of law, especially unwritten law, advised the praetor on the content of his edict, and assisted parties and judges in litigation. Augustus empowered certain jurists to give *responsa* with the emperor's authority; this increased their prestige, but the practice lapsed as early as AD 200.

During the early empire, numerous commentaries were written by the great jurists on individual *leges,* on civil law, on the edict, and on law as a whole. In the 5th century a law was passed stipulating that only the works of certain jurists could be cited. Legal scholarship declined in the postclassical period.

The law of Justinian. When the Byzantine emperor Justinian I assumed rule in AD 527, he found the law of the Roman Empire in a state of great confusion. It consisted of two masses that were usually distinguished as old law and new law.

Old law and new law under Justinian

The old law comprised (1) all of the statutes passed under the republic and early empire that had not become obsolete; (2) the decrees of the Senate passed at the end of the republic and during the first two centuries of the empire; and (3) the writings of jurists and, more particularly, of those jurists to whom the emperors had given the right of declaring the law with their authority. These jurists, in their commentaries, had incorporated practically all that was of importance. Of these numerous records and writings of old law, many had become scarce or had been lost altogether, and some were of doubtful authenticity. The entire mass of work was so costly to produce that even the public libraries did not contain complete collections. Moreover, these writings contained many inconsistencies.

The new law, which consisted of the ordinances of the emperors promulgated during the middle and later stages of the empire, was in a similarly disorganized condition. These ordinances or constitutions were extremely numerous and contradictory. Because no complete collection existed (earlier codices were not comprehensive), other ordinances had to be obtained separately. It was thus necessary to collect into a reasonable corpus as much of the law, both new and old, as was regarded as binding and to purge its contradictions and inconsistencies.

Immediately after his accession, Justinian appointed a commission to deal with the imperial constitutions. The 10 commissioners went through all of the constitutions of which copies existed, selected those that had practical value, cut all unnecessary matter, eliminated contradictions by omitting one or the other of the conflicting passages, and adapted all the provisions to the circumstances of Justinian's own time. The resulting Codex Constitutionum was formally promulgated in 529, and all imperial ordinances not included in it were repealed. This Codex has been lost, but a revised edition of 534 exists as part of the so-called Corpus Juris Civilis.

The success of this first experiment encouraged the Emperor to attempt the more difficult enterprise of simplifying and digesting the writings of the jurists. Thus, beginning in 530, a new commission of 16 eminent lawyers set about this task of compiling, clarifying, simplifying, and ordering; the results were published in 533 in 50 books that became known as the Digest (Digesta) or Pandects (Pandectae). After enacting the Digest as a lawbook, Justinian repealed all of the other law contained in the treatises of the jurists and directed that those treatises should never be cited in the future, even by way of illustration; at the same time, he abrogated all of the statutes that had formed a part of the old law. An outline of the elements of Roman law called the Institutes of Justinian (or simply Institutiones) was published at about the same time.

Between 534 and his death in 565, Justinian himself issued a great number of ordinances that dealt with many subjects and seriously altered the law on many points. These ordinances are called, by way of distinction, new constitutions (Novellae Constitutiones Post Codicem); in English they are referred to as the Novels.

Corpus Juris Civilis: the civil-law code of Justinian

All of these books—the revised Codex Constitutionum (the original work was revised four and a half years later), the Digest, the Institutes, and the Novels—are collectively known as the Corpus Juris Civilis. This Corpus Juris of Justinian, with a few additions from the ordinances of succeeding emperors, continued to be the chief lawbook in what remained of the Roman world. In the 9th century a new system known as the Basilica was prepared by the emperor Leo VI the Wise. It was written in Greek and consisted of parts of the Codex and parts of the Digest, joined and often altered in expression, together with some material from the Novels and imperial ordinances subsequent to those of Justinian. In the western provinces, the law as settled by Justinian held its ground.

Categories of Roman law. *The law of persons.* "The main distinction in the law of persons," said the 2nd-century jurist Gaius, "is that all men are either free or slaves." The slave was, in principle, a human chattel who could be owned and dealt with like any other piece of property. As such, he was not only at the mercy of his owner but rightless and (apart from criminal law) dutiless. Even though the slave was in law a thing, he was in fact a man, and this modified the principle. A slave could not be a party to a contract nor own property, but he could be given a de facto patrimony, which could be retained if he were freed; if he made a "commitment," it could ultimately be enforced against his master. A manumitted slave became, in most instances, not only free but also a citizen.

The definition of citizenship was important for the purposes of private law because certain parts applied only to citizens (*jus civile*). Noncitizens could be either Latini, inhabitants of Roman settlements that had the rights of members of the original Latin League, or *peregrini,* who were members of foreign communities or of those territories governed but not absorbed by Rome. The great extension of the citizenship by the emperor Caracalla in AD 212 reduced the importance of this part of the law.

Patria potestas: the power of the paternal head of the Roman family

Family. The chief characteristic of the Roman family was the *patria potestas* (paternal power in the form of absolute authority), which the elder father exercised over his children and over his more remote descendants in the male line, whatever their age might be, as well as over those who were brought into the family by adoption—a common practice at Rome. Originally this meant not only that he had control over his children, even to the right of inflicting capital punishment, but that he alone had any rights in private law. Thus, any acquisitions made by a child under *potestas* became the property of the father. The father might indeed allow a child (as he might a slave) certain property to treat as his own, but in the eye of the law it continued to belong to the father.

By the 1st century AD there were already modifications of the system: the father's power of life and death had shrunk to that of light chastisement, and the son could bind his father by contract with a third party within the same strict limits that applied to slaves and their masters. Sons also could keep as their own what they earned as soldiers and even make wills of it. In Justinian's day, the position regarding property had changed considerably. What the father gave to the son still remained, in law, the father's property, but the rules concerning the son's own earnings had been extended to many sorts of professional earnings; and in other acquisitions (such as property inherited from the mother), the father's rights were reduced to a life interest (usufruct). Normally, *patria potestas* ceased only with the death of the father; but the father might voluntarily free the child by emancipation, and a daughter ceased to be under her father's *potestas* if she came under the *manus* of her husband.

Marriage with and without *manus*, or husbandly authority

There were two types of marriage known to the law, one with *manus* and one without, but the *manus* type of marriage was rare even in the late republic and had disappeared long before Justinian's day. *Manus* was the autocratic power of the husband over the wife, corresponding to *patria potestas* over the sons.

Marriage without *manus* was by far the more common in all properly attested periods. It was formed (provided the parties were above the age of puberty and, if under *potestas*, had their father's consent) simply by beginning conjugal life with the intention of being married, normally evidenced by the bringing of the bride to the bridegroom's house. The wife remained under her father's *potestas* if he were still alive; if he were dead, she continued (as long as guardianship of women continued) to have the same guardian as before marriage. Both spouses had to be citizens, or if one was not, he or she must have *conubium* (the right, sometimes given to non-Romans, of contracting a Roman marriage). In marriage without *manus*, the property of the spouses remained distinct, and even gifts between husband and wife were invalid.

Divorce was always possible at the instance of the husband in cases of marriage with *manus;* in marriage without *manus*, either party was free to put an end to the relationship. A formal letter was usually given to the spouse, but any manifestation of intention to end the relationship—made clear to the other party and accompanied by actual parting—was all that was legally necessary. The Christian emperors imposed penalties on those who divorced without good reason, but the power of the parties to end the marriage by their own act was not taken away.

Concubinage was recognized in the empire as a "marriage" without a dowry, with a lower status for the woman, and with provisions that the children were not legally the father's heirs. A man could not have both a wife and a concubine. In the 4th century the emperor Constantine first enacted a law enabling the children of such unions to be legitimated by the subsequent marriage of their parents. Medieval civil law extended this rule to all illegitimate children.

Guardianship

Persons under the age of puberty (14 for males, 12 for females) needed *tutores* if they were not under *patria potestas*. Such tutors could be appointed under the will of the father or male head of the household. Failing such an appointment, the guardianship went to certain prescribed relatives; if there were no qualified relations, the magistrates appointed a tutor. Originally, children were considered adults at the age of puberty; but, after a long development, it became usual for those between the ages of puberty and 25 to have guardians who were always magisterially appointed. Originally, all women not under *patria potestas* or *manus* also needed *tutores*, appointed in the same way as those for children. By the early empire, this provision was little more than a burdensome technicality, and it disappeared from Justinian's law.

Corporations. The Romans did not develop a generalized concept of juristic personality in the sense of an entity that had rights and duties. They had no terms for a corporation or a legal person. But they did endow certain aggregations of persons with particular powers and capacities, and the underlying legal notion hovered between corporate powers, as understood in modern law, and powers enjoyed collectively by a group of individuals. The source of such collective powers, however, was always an act of state.

Four types of corporation were distinguished:

1. *Municipia* (the citizen body, originally composed of the conquered cities and later of other local communities) possessed a corporateness that was recognized in such matters as having the power to acquire things and to contract. In imperial times, they were accorded the power to manumit slaves, take legacies, and finally—though this became general only in postclassical law—to be instituted as an heir.

2. The *populus Romanus*, or the "people of Rome," collectively could acquire property, make contracts, and be appointed heir. Public property included the property of the treasury.

3. *Collegia*—numerous private associations with specialized functions, such as craft or trade guilds, burial societies, and societies dedicated to special religious worship—seem to have carried on their affairs and to have held property corporately in republican times. The emperors, viewing the *collegia* with some suspicion, enacted from the beginning that no *collegium* could be founded without state authority and that their rights of manumitting slaves and taking legacies be closely regulated.

4. Charitable funds became a concern of postclassical law. Property might be donated or willed—normally, but not necessarily, to a church—for some charitable use, and the church would then (or so it appears from the evidence) have the duty of supervising the fund. Imperial legislation controlled the disposition of such funds so that they could not be used illegally. In such cases ownership is thought to have been temporarily vested in the administrators.

The Roman concept of absolute ownership

The law of property and possession. In Roman law (today as well as in Roman times), both land and movable property could be owned absolutely by individuals. This conception of absolute ownership (*dominium*) is characteristically Roman, as opposed to the relative idea of ownership as the better right to possession that underlies the Germanic systems and English law.

Mancipatio, or formal transfer of property, involved a ceremonial conveyance needing for its accomplishment the presence of the transferor and transferee, five witnesses (adult Roman citizens), a pair of scales, a man to hold them, and an ingot of copper. The transferee grasped the object being transferred and said, "I assert that this thing is mine by Quiritarian [Roman] law; and be it bought to me with this piece of copper and these copper scales." He then struck the scales with the ingot, which he handed to the transferor "by way of price."

In jure cessio was a conveyance in the form of a lawsuit. The transferee claimed before the magistrate that the thing was his, and the transferor, who was the defendant, admitted the claim. The magistrate then adjudged the thing to the transferee. (The sham-lawsuit theory, however, is not acceptable to all modern scholars, principally because the judgment of ownership was valid against any possible private claimant, not merely against the defendant, as in a true lawsuit.)

Usucapio referred to ownership acquired by length of possession. In early Roman law, two years of continuous possession established title in the case of land, one year in the case of movables. In the developed law, possession must have begun justifiably in good faith, and the thing must not have been stolen (even though the possessor himself may have been innocent of the theft) or acquired by violence.

In terms of *occupatio*, ownerless things that were susceptible to private ownership (excluding such things as temples) became the property of the first person to take possession of them. This applied to things such as wild animals and islands arising in the sea. In some views, it also applied to abandoned articles.

Accessio worked in this manner: if an accessory thing belonging to A was joined to a principal one belonging to B, the ownership in the whole went to B. For example, if A's purple were used to dye B's cloth, the dyed cloth belonged wholly to B. By far the most important application of this rule asserted that whatever is built on land becomes part of the land and cannot be separately owned.

Specificatio was somewhat different. If A made a thing out of material belonging to B, one school of thought held that ownership went to A, and another held that it remained with B. Justinian adopted a "middle opinion": B retained ownership if reconversion to the original condition was possible (a bronze vase could be melted down); A obtained ownership if it was not (wine cannot be reconverted into grapes).

According to *thesauri inventio*, or treasure trove, the final rule was that if something was found by a man on his own land, it went to him; if it was found on the land of another, half went to the finder, half to the landowner.

Traditio was the simple delivery of possession with the intention of passing ownership and was the method of conveyance of the *jus gentium*. If A sold and merely delivered a slave to B, under the *jus civile*, A remained the owner of the slave until a specified length of time had elapsed. The praetors, however, devised procedural methods of protecting B's possession in such a way that A's title became valueless, and B was said to own the thing *in bonis*. This was a remarkable triumph for informality in the granting of title. From the phrase *in bonis*, later writers coined the expression "bonitary ownership." Justinian abolished the theoretical distinction between civil and bonitary ownership.

The ordinary leaseholder had no protection beyond a contractual right against a landlord and could not assign tenancy. But there were certain kinds of tenure that did provide the tenant protection and that were assignable: agricultural and building leases granted for a long term or in perpetuity often enabled leaseholders to enjoy rights hardly distinguishable from ownership.

Servitudes: rights to specific, limited use or enjoyment of another's property

There were also servitudes, in which one person enjoyed certain rights in property owned by another. Rights of way and water rights were rustic servitudes; rights to light or to view were urban servitudes. *Ususfructus* was the right to use and take the fruits (such as crops) of a thing and corresponded to the modern notion of life interest. A more restricted right, likewise not extending beyond the life of the holder, *usus* permitted merely the use of a thing; thus, a person could live in a house but could not let it, as that would be equivalent to "taking the fruits."

Since ownership was absolute, it was sharply distinguished from possession, which the civil law did not protect as such. Any owner wishing to interfere with an existing possessor, however, had to bring legal action to prove his title. If he interfered on his own authority, the praetor would see that the original state of affairs was restored before adjudicating the title.

Delict and contract. Obligations were classified by classical jurists into two main categories, according to whether they arose from delict or contract. Justinian's law recognized two further classes of obligation, termed quasi-delict and quasi-contract.

As early as the 6th and 5th centuries BC, Roman law was experiencing a transition from a system of private vengeance to one in which the state insisted that the person wronged accept compensation instead of vengeance. Thus, in the case of assault (*injuria*), if one man broke another's limb, *talio* was still permitted (that is, the person wronged could inflict the same injury as he had received); but in other cases, fixed monetary penalties were set. Theft involved a penalty of twice the value of the thing stolen, unless the thief was caught in the act, in which case he was flogged and "adjudged" to the person wronged.

By the early empire, reforms had substituted a fourfold penalty in the case of a thief who was caught in the act, and the court assessed all penalties for *injuria* (which by then included defamation and insulting behaviour). The law of damage to property was regulated by statute (the Lex Aquilia), which in turn was much extended by interpretation. Additionally, there were situations in which a person could be held liable for damages even though he was not personally responsible.

In the early republic, a law of contract hardly existed. There was, however, an institution called *nexum*, of which little can be said with certainty except that it was a kind of loan so oppressive in character that it could result in the debtor's complete subjection to the creditor. It was obsolete long before imperial times. The contracts of classical law were divided into four classes: literal, verbal, real, and consensual. The literal contract was a type of fictitious loan formed by an entry in the creditor's account book; it was comparatively unimportant and was obsolete by Justinian's day. The verbal contract, or *stipulatio*, was of great importance, for it established a form in which any agreement (provided it was lawful and possible) could be made binding by the simple method of reducing it to question and answer: "Do you promise to pay me 10,000 sesterces?" "I promise." Originally it was absolutely necessary that the words be spoken, but by Justinian's day a written memorandum of such a contract would be binding, even though, in fact, nothing at all had been spoken.

If an agreement was not clothed in the form of a stipulation, it must, to be valid, fall under one of the types of real or consensual contracts. A real contract was one requiring that something should be transferred from one party to the other and that the obligation arising should be for the return of that thing. Real contracts included loans of money, loans of goods, deposits, and pledges. Consensual contracts needed nothing except verbal or written agreement between the parties, and though there were only four such contracts known to the law, they were the most important in ordinary life—sale, hire of things or services, partnership, and mandate (acting upon instructions). In Justinian's day there was a further principle that in any case of reciprocal agreement, such as an agreement for exchange (but not sale), if one party had performed, he could bring an action to enforce performance by the other. In addition to the foregoing contracts, a few other specific agreements were recognized as enforceable, but the general recognition of all serious agreements as binding was never achieved by the Romans.

Enforceable contracts

Quasi-delict covered four types of harm, grouped together by no clearly ascertainable principle. They included the action against an occupier for harm done by things thrown or poured from his house into a public place and the action against a shipowner, innkeeper, or stablekeeper for loss caused to customers on the premises through theft or damage by persons in his service.

Quasi-contract embraced obligations that had no common feature save that they did not properly fall under contract, because there was no agreement, or under delict, because there was no wrongful act. The most noticeable examples were, first, *negotiorum gestio*, which enabled one who intervened without authority in another's affairs for the latter's benefit to claim reimbursement and indemnity, and second, the group of cases in which an action (*condictio*) was allowed for the recovery by A from B of what would otherwise be an unjustified enrichment of B at A's expense, such as when A had mistakenly paid B something that was not due (*condictio indebiti*). This notion of unjust enrichment as a source of legal obligation was one of the most pregnant contributions made by Roman law to legal thought.

The law of succession. The first requirement of any Roman will of historical times was the appointment of one or more heirs. An heir, in the Roman sense of the term, was a universal successor; that is, he took over the rights and duties of the deceased (insofar as they were transmissible at all) as a whole. On acceptance, the heir became owner if the deceased was owner, creditor if he was creditor, and debtor if he was debtor, even though the assets were insufficient to pay the debts. It was thus possible for an inheritance to involve the heir in a loss. Until Justinian's day this consequence could be avoided only by not accepting the inheritance, but Justinian made one of his most famous reforms by providing that an heir who made an inventory of the deceased's assets need not pay out more than he had received. Freedom of testation, furthermore, was not complete: a man was obliged to leave a certain proportion of his property to his children and in some cases to ascendants and brothers and sisters.

With regard to intestate succession, or succession without a will, those first entitled in early times were the deceased's own heirs—that is, those who were in his *potestas* or *manus* when he died and who were freed from that power at his death. Failing these heirs, the nearest agnatic relations (relations in the male line of descent) succeeded, and, if there were no agnates, the members of the gens, or clan, of the deceased succeeded. Later reforms placed children emancipated from *potestas* on an equal basis with those under *potestas* and gradually gave the surviving spouse (in marriage without *manus*) greater rights of succession. By Justinian's day the system had evolved as follows: descendants had the first claim, and failing these heirs, came a composite class consisting of ascendants, brothers and sisters of full blood, and children of deceased brothers and sisters. Next came brothers and sisters of the half blood and, finally, the nearest cognates (relations in the female line). Husband and wife were not mentioned, but their old rights were kept alive in the absence of any of the preceding categories. Justinian also gave a "poor" widow a right to one-quarter of her husband's estate unless there were more than three children, in which case she shared equally with them. If, however, the heirs were her own children by the deceased, she received only a *ususfructus* (life interest) in what she took.

The law of procedure. The earliest forms of procedure had two stages: a preliminary one before the jurisdictional magistrate, in which the issue was developed; and then the actual trial before the *judex*, or judge. The system required that set forms of words be spoken by the parties and, sometimes, by the magistrate. The parties making an assertion of ownership, for instance, would grasp the thing in dispute and lay a wand on it, after which the magistrate would intervene and say, "Let go, both of you." So formal was the procedure that a plaintiff who made the slightest mistake lost his case. Under the system the plaintiff was also responsible for physically producing the defendant in court and, often, for carrying out the sentence of the court.

Instructions to the judge

Under new procedures developed in the 2nd and 1st centuries BC, the issue at the magisterial stage was formulated in written instructions to the *judex*, couched in the form of an alternative: "If it appears that the defendant owes the plaintiff 10,000 sesterces, the *judex* is to condemn the defendant to pay the plaintiff 10,000 sesterces; if it does not so appear, he is to absolve him." A draft of these written instructions was probably prepared for the plaintiff before he came into court, but there could be no trial until it was accepted by the defendant, for there was always a contractual element about a lawsuit under both the new and the old systems. Pressure, however, could be exercised by the magistrate on a defendant who refused to accept instructions that the magistrate had approved, just as a plaintiff could be forced to alter instructions that the magistrate had disapproved, by the magistrate's refusal to otherwise give the order to the *judex* to decide the case.

In late republican times, still another system developed, first in the provinces, then in Rome. Under the new system the magistrate used his administrative powers, which were always considerable, for the purpose of settling disputes. He could command: thus if one person brought a complaint against another before him, he could investigate the matter and give the order he thought fit. As imperially appointed officers superseded republican magistrates, this administrative process became more common. The result was that the old contractual element in procedure disappeared as did the old two-stage division. Justice was now imposed from above by the state—not, as originally, left to a kind of voluntary arbitration supervised by the state.

(H.F.J./R.Po./M.A.M./M.A.Gl.)

GERMANIC LAW

Germanic law is a designation that covers the laws of the various peoples of Germanic stock from the time that the earliest "barbarian" tribes came into contact with the Romans until their tribal laws developed into national territorial laws—a development that occurred at different times with different peoples. Thus some of the characteristics of Scandinavian legal collections of the 12th century are similar to those in the Visigothic laws of the 6th century.

Knowledge of the early Germanic period is derived mainly from the observations of tribal life contained in Julius Caesar's *Gallic War* and Tacitus' *Germania*. The first written collections of Germanic law are the so-called Leges Barbarorum, which date from the 5th century until the 9th century. They are written in Latin and show Roman influence by their use of the technical terms of Roman law. The Anglo-Saxon laws and the laws of the North Germanic group, on the other hand, are in the vernacular and owe their written form largely to the advent of Christianity.

For all of the Germanic peoples, law (West German, *reht* and *êwa;* High German, *wizzôd;* North German, *lagh*, from which the English word *law* is derived) was basically not something laid down by a central authority, such as the king, but rather the custom of a particular nation (tribe). It was essentially unwritten, being derived from popular practices, and was not sharply distinguished from morality; it was personal in the sense that it applied only to those who belonged to the nation. Thus each man followed his own law, a notion appropriate to a nomadic people who originally did not live in a clearly defined territory. When, after the fall of the Roman Empire in the West, Germanic tribes took over former Roman provinces, they did not attempt to apply their laws to their Roman subjects, for whom Roman law remained applicable.

Thus the earliest Germanic code, that of Euric, king of the Visigoths in Spain and southwestern Gaul in the late 5th century, applied exclusively to Visigoths. The Lex Romana Visigothorum, or Breviary of Alaric, was issued in AD 506 for their Roman subjects. It was a compilation of "vulgar law"—Roman law adapted to fit the social and economic conditions of the late Roman Empire—and was later the main source of Roman law in the Frankish kingdom. Only in the 7th century was Visigothic law applied to Visigoths and Romans alike, the two peoples by then having substantially fused. The Lex Burgundiorum and the Lex Romana Burgundiorum of the same period had similar functions, while the Edictum Rothari (643) applied to Lombards only.

Visigothic laws

The Leges Barbarorum, then, were not legislation in the modern sense but rather the records of customs that were first collected and then declared as law. The prologue to the Salic Law (the law of the West, or Salic, Franks) recounted how four chosen men collected the original practices in particular cases, having first discussed them with the presidents of the local popular assemblies. The Leges Barbarorum did not seek to set out all of the main rules of law as modern codes do. They were not concerned with what everyone took for granted but concentrated on matters that, perhaps as a result of migration or conquest, had become doubtful and needed authoritative exposition. They dealt with specific situations rather than general rules and focused particularly on court procedure, monetary compensation for acts of violence, and succession on death.

The initiative for declaring law usually came from the king, but the resulting laws normally required approval by the popular assemblies. Because of this collaboration between king and people, a compilation was sometimes referred to as an "agreement," or *pactus*. The Visigothic laws were an exception; they always appear to have been formulated by the king and chief landowners without popular participation. Gradually, first the Lombard and then the Frankish kings overcame their people's aversion to central government and began to legislate unilaterally. The Lombards, who invaded Italy in 568, had no single code of custom, but their kings issued edicts from the mid-7th century onward. In the Frankish kingdom the Merovingian kings called their legislation *edicta* or *praecepta*, but the succeeding Carolingians characterized them as *capitularia; i.e.*, royal ordinances divided into articles (*capitula*). These included modifications of the *leges* of the Franks or other nations in the Frankish kingdom, administrative orders to officials, and independent legislation. Like the Roman emperors before them, Charlemagne and his successors claimed the power to make laws for all their subjects, irrespective of nation, and without the consent of any assembly. The validity of the law depended solely on the oral act of the king who promulgated it.

Tribal Germanic institutions. Germanic law recognized a distinction between free and unfree persons. Only the former had legal capacity, and they were subdivided into nobles and ordinary freemen. The nobles enjoyed a larger share in land distribution, were preferentially chosen for public office, and were protected by a larger monetary compensation if they were injured. Certain west Germanic tribes recognized an intermediate status of half-free persons, who could enter into legal transactions and marry but had no political rights.

Basically a Germanic tribe was a league of clans. Its main institutions of government were the king, his council, and the tribal assembly (*mallus, witan, mot, ding*, or *thing*). The king was military leader, chief priest, and president of the assembly, and he was assisted in the routine business of government by his council of elders and higher nobles. The assembly was composed of all free members of the

Role of the king

tribe grouped into clans. It elected kings, declared war, outlawed freemen, and generally controlled the membership of the tribe by its supervision of the manumission of slaves, the emancipation of minors, and the adoption of strangers.

The dominant social institution was the "sib" (*sippe*), a term that meant both a clan—the extended family composed of all those related by blood, however remotely, and subject to a clan chief—and also a household or narrow family, whose members were under the *mund* (guardianship) of the family head. A boy remained in his father's *mund* until he was emancipated on attaining physical maturity; a girl remained until she married, when she passed into the *mund* of her husband. Marriage commonly took the form of the sale of the bride to her groom for a price, which developed into a fund held by the husband for the wife's benefit. A husband could divorce his wife at will but risked being penalized financially.

Laws of property

The main notion in the law of property was *gewere*, or the power exercised by the owner, which did not clearly distinguish between legal title and physical control. Various forms of limited ownership were recognized. Land was treated differently from movables; originally it had belonged to each family collectively. Family ownership gradually developed into the private ownership of the family head, but for a long time he could alienate land only with the consent of the nearest heirs. Land transfer required much formality, and among the west Germanic peoples a glove or spear was handed over as a symbol of the transfer of *gewere*.

At the death of the family head, his property passed to his descendants in the nearest degree of proximity, with a preference for males. (The declaration in the Salic Law that daughters could not inherit land was used by 16th-century French lawyers as additional support for the long-standing practice of excluding women or their descendants from succeeding to the crown.) In the absence of descendants, several *leges* provided that property deriving from the father's side should return to that side and property from the mother's side to her side. The order of succession could not be altered by will.

When trade was still conducted on a cash or barter basis, there was little need for formal contract law. A family could obligate itself to another either by pledging a thing as security (*wadium, gage*) or by surrendering a hostage (*gijzel, born*).

Later, a debt was guaranteed by a formal oath accompanied by the surrendering of a staff to the creditor (*effestucatio*). Contractual obligation was then constituted either by oath (enforced by an action for perjury) or by delivery of a thing (enforced by an action for theft).

Laws to control crime

Offenses against the community, such as treason, secret killing, and secret theft, were punished by outlawry, which was pronounced by the tribal assembly. The convicted person could then be killed by anyone. Offenses against individuals, including open killing and open robbery, became the subject of a blood feud if the criminal and victim belonged to different family groups. Peace could be bought by the payment of compensation, known as *wergild* in homicide cases and *bot* in others. Payment was voluntary at first; only later did it become obligatory. Even in the 7th century, Visigothic law still allowed retaliation in kind for all injuries except those to the head. The *leges* contained elaborate tariffs of compensation for different kinds of injury, the amount varying according to the social status of the victim. Private feuds were eventually restricted by the growth of royal authority in the Frankish period and the notion of the king's peace, the breach of which was punishable by the king's court.

When parties appeared before a court and stated their cases, the court decided on an acceptable method of proof, which could be by oath of the parties, supported by *compurgatores* (literally "oath-helpers"), the number required depending on the gravity of the case, by ordeal, or by battle. A successful claimant had to enforce judgment himself on the person or property of the defendant.

Rise of feudal and monarchial states. With the disintegration of the Frankish kingdom in the late 9th century, government became highly decentralized. Already the pattern of landholding, which determined the more important legal relationships, had begun to take on the characteristics of feudalism. Before the end of the Roman Empire much of the land had been concentrated in the hands of magnates, secular and ecclesiastical. But, unlike their predecessors under the Romans, the holders of secular land in the Germanic states became largely independent of the central government. By the 9th century, many lords had become strong enough to challenge the power of the Carolingian kings of the Frankish Empire and to make the inhabitants of their own areas their vassals. These vassals held their land from the lords as tenants of a so-called feud, or fee. Each feudal lord held a court for his tenants in which he applied the same law to all of the tenants, irrespective of their racial or national origin. Thus the old Germanic personal principle was abandoned in favour of the territorial principle, or the application of the custom of the region. This type of feudal law usually was based partly on Germanic law and partly on the Roman law of the Lex Romana Visigothorum, adapted in the interests of the feudal lords.

Influence of canon law

During the same period the Roman Catholic Church became the main unifying force in western Europe and began to claim jurisdiction over many matters that earlier had been considered secular rather than ecclesiastical. Church courts had existed since the Roman Empire, and their power in matters of faith was recognized by the secular authorities. The personal law of the church as an institution was always Roman, and indeed the law of the Ripuarian Franks on the Rhine expressly declared that "the church lives by Roman law." The canon law applied in the church courts was largely influenced by Roman law and contained very few Germanic elements. As the power of the church grew, the church courts applied this law to matters that had previously been dealt with by the secular courts, such as marriage, adultery, wills, and succession. In many countries these matters remained withdrawn from Germanic law and subject to church law even after the Reformation.

Merchants also found that the old Germanic customary law was inadequate to cope with the problems created by the rapid growth of commerce that had occurred by the 12th century. A special commercial law, based mainly on Roman law as developed by the Mediterranean seaborne traders, was developed to settle disputes between merchants, without regard to their nationality or place of residence.

These developments reduced the range of cases that were subject to the jurisdiction of the local county courts. In Germany some of the earlier codifications of customary law were forgotten, partly because the local judges were unable to understand the Latin in which they were written and partly because the rules that they contained were unsuited to the new social and economic conditions. The local courts applied an unwritten customary law based on the dominant tribal law of the area, and it was this that formed the basis of such codifications as the *Sachsenspiegel* ("Mirror of the Saxons") in the 13th century.

Regional variations

In France the legal development in the north differed from that in the south. The regional customs in the north were made up of Germanic and Roman law, the Carolingian capitularies, and canon law, but Germanic elements predominated. In the south, the so-called *pays de droit écrit* ("land of written law"), where Gallo-Romans had been far more numerous than Franks, the custom of each district was based mainly on the vulgar law of the Lex Romana Visigothorum. In Italy this law existed side by side with Lombard law. In the 7th and 8th centuries that law was subjected to a relatively sophisticated codification, whose form showed Roman influence.

In England the Norman conquerors continued the movement toward legal unity begun by the Anglo-Saxons by imposing on the country a centralized form of government more powerful than any on the Continent. In the 12th century Henry II made the king's court a permanent court of professional judges with jurisdiction over many matters that earlier had been dealt with by other courts. The common law developed by this court was largely Germanic law.

(P.G.S./M.A.Gl.)

Civil-law systems

THE HISTORICAL RISE OF CIVIL LAW

In the 5th and 6th centuries western and central Europe were dominated by Germanic peoples, especially those who had overrun the Roman Empire. Among them were the Anglo-Saxons of England, the Franks of western Germany and northern France, the Burgundians, the Visigoths of southern France and Spain, and the Lombards of Italy. Although Roman law traditions lingered on for some time, the Germanic customs came to prevail in most regions. In the Middle Ages these customs underwent vigorous growth in an effort to satisfy the complex needs of a society that encompassed changing feudalism, chivalry, growing cities, Eastern colonization, increasing trade, and a constantly refined culture. Among the many strands that went into the weaving of the complex pattern of medieval law, the customs of the merchants and the canon law of the Roman Catholic Church were of special significance. It was through the canon law that the concepts and ideas of ancient Rome continued to make their presence felt, even when, as a whole, Roman law had been forgotten. In the late 11th century Roman law was rediscovered and made the subject matter of learned study and teaching by scholars in northern Italy, especially at Bologna. With the increasing demand for trained judges and administrators, first by the Italian city-republics, then by princes in other localities, students flocked to Bologna from all over Europe, until the study and teaching of law were gradually taken over by local universities. As a result of this process, Roman law penetrated into the administration of justice north of the Alps, especially in Germany and the Netherlands, where the Roman-law influence became particularly strong.

Revival of Roman law

In the Holy Roman Empire of the German Nation the reception of Roman law was facilitated because its emperors cherished the idea of being the direct successors of the Roman Caesars; Roman law, collected in the Corpus Juris Civilis by the emperor Justinian I between 527 and 565, could be regarded as still being in effect simply because it was the imperial law. Decisive for the reception, however, was the superiority of the specialized training of Roman-law jurists over the empiricist methods of lay judges and practitioners of the local laws. Equally decisive was the superiority of the Roman-canonical type of procedure, with its rational rules of evidence, over the local forms of procedure involving proof by ordeal, battle, and other irrational methods. Nowhere, however, did Roman law completely supplant the local laws, and, as far as the content of the law was concerned, various amalgams developed. Roman law strongly influenced the law of contracts and torts; canon law achieved supremacy in the field of marriage; and combinations of Germanic, feudal, and Roman traditions developed in matters of property and succession. The conceptual formulations in which the norms and principles of the law were expressed, as well as the procedural forms in which justice was administered, were also strongly Roman. The system that thus emerged was called the *jus commune.* In actual practice it varied from place to place, but it was nevertheless a unit that was held together by a common tradition and a common stock of learning. Although the law of the Corpus Juris Civilis (especially its main part, the Digest, the writings of the jurists) was, as such, in effect nowhere, it constituted the basis of study, training, and discourse everywhere. In spite of all local variety, the civil-law world experienced a sense of unity that corresponded to the strongly felt unity of European civilization.

This unity was undermined by the religious split of the Reformation and Counter-Reformation and by the rise of nationalism that accompanied the unification and stabilization of the European nations and their struggle for hegemony. In the field of law the split found expression in the national codifications, through which the law was unified within each nation but simultaneously was set apart from that of all others. In Denmark codification occurred in 1683, in Norway in 1687, in Sweden-Finland in 1734, and in Prussia in 1794. Because of the personality of their promoter and the novel technique applied, great fame and influence were achieved by the Napoleonic codifications of the private and criminal law of France, especially their central piece, the Civil, or Napoleonic, Code (Code Civil or Code Napoléon) of 1804.

National codifications

Codification continued after the Napoleonic era. In Belgium and Luxembourg, which had been incorporated into France under Napoleon, his codes were simply left in effect. The Netherlands, Italy, Spain, Portugal, and numerous countries of Latin America followed the French model not only by undertaking national codification but also by using the same techniques and arrangements. Naturally, their courts and legal scholars were, at least in the early 19th century, inclined to pay great attention to French legal learning.

In Germany national codification came considerably later than in France. Only a commercial code had been uniformly created by the independent German states shortly after the revolution of 1848. The unification of the criminal law took place almost simultaneously with the political unification of the country, which occurred in 1871. Codification of the organization of the courts and of civil and criminal procedure came in 1879. But the Civil Code (Bürgerliches Gesetzbuch für das deutsche Reich) was not completed until 1896, and it did not take effect until Jan. 1, 1900.

Throughout the 19th century the vigorous German science of law exercised much influence in Austria (which as early as 1811 had codified its law in a technique different from that of France), in Switzerland, in the Nordic countries, and, later, in most of eastern Europe. When Swiss law was codified in 1907–12, it became the model for the Turkish codification of 1926 and strongly influenced that of China, which is still in effect in Taiwan.

Due to the different dates of codification and the different style and attitude of legal learning, the civil-law family of laws is thus divided into the French, or Romanist, branch and the German, or Germanic, branch. Their main features are determined by those of their prototypes. The legal system of Japan essentially belongs to the German branch, but it presents important features of its own.

THE FRENCH SYSTEM

Changes induced by the French Revolution

In France the revolutionary period was one of extensive legislative activity, and long-desired changes were enthusiastically introduced. A new conception of law appeared in France: statute was deemed the basic source of law. Customs remained only if they could not be replaced by statutes. The Parlements, the major courts of the nation, were dismantled and replaced by a unified system of courts that were merely supposed to apply the law and never to lay down general rules.

The main ideas embodied in the revolutionary legislation were to be found in the motto of the Revolution (which is still that of France), "*Liberté, égalité, fraternité.*" The passionate desire for liberty and equality aroused by the 18th-century philosophers inspired the changes that took place.

The system that had come to be called feudal, although it had little to do with the feudalism of the High Middle Ages, was hated by the peasants and the bourgeoisie for its unbalanced distribution of privileges—especially those exempting the nobles and clergy from taxation. These privileges were abolished early in the Revolution. The revolutionaries detested organized groups of any kind, for it was thought that only one authority should exist over the citizens—that of the state. As a result the guilds, which demanded compulsory membership and regulated every profession, were suppressed, and freedom of commerce was established. The old-style universities were dissolved; in the same spirit the property of the Roman Catholic Church was secularized, and the priests and bishops were made state employees, a situation that most of them did not accept.

Family relations were deeply transformed according to the principles of liberty and equality. Marriage was organized merely as a civil act, divorce was permitted, paternal authority was limited, and parents' consent was not required for marriages of children over 21 years of age. A short experiment was made with "family courts" that were permitted to overrule paternal decisions, and the wife was declared equal to her husband. In matters of succession, equal parts were given to all children, and the testator's right to dispose of property by will was limited in order to prevent the reestablishment of inequalities by this device.

Throughout the revolutionary period, successive governments were committed to consolidating the legal changes in a set of codes. Drafts were made, but time and authority were lacking, and none was enacted until society was restabilized under Napoleon.

The concept of codification. From a practical point of view, the Civil Code achieved the unification of French civil law. This was not, however, the only concern of its drafters. They shared with most of their contemporaries and with most modern French lawyers the belief that the law should be written in clear language so that it would be accessible to every citizen. This view implied that the new code must be complete in its field, setting forth general rules and arranging them logically. Finally, it must not unnecessarily break with tradition.

The Civil Code was organized as a series of short articles because it was assumed, first, that legislators could not foresee all circumstances that might arise in life and, second, that only conciseness could make the code flexible enough to adapt old principles to new circumstances. The general rules contained in the code have since been applied to concrete circumstances without much difficulty. When an interpretation has been required, the courts have had the responsibility to give it, taking into consideration the "spirit" of the code in an effort to apply to each case the solution that would have been desired by the legislator.

Logic and experience in the codification

The drafters of the code strove toward inner consistency in their work, so that reliance on logic might ensure satisfactory application of it. They saw no contradiction between logic and experience. Since the 17th-century beginnings of the Age of Reason, abstract reasoning had characterized the French approach to law and to life in general. For this reason articles of the code were not regarded as narrow rulings. If no one article was found to apply exactly to a given situation, it was proper to consider several articles and to draw from them a more general rule that could either itself be applied to the case or be combined with others to reach a solution.

Although the code was a work of logic, it relied mainly on experience. Its drafters were exceptionally well qualified in this respect: they had lived the first half of their lives under the laws of ancient France and had also known the Revolution. Their purpose was not so much to create new laws as to restate existing laws, subject to choice when revolutionary enactments varied from previous ones and when previous laws differed from one another. They were ready to adopt any rules that seemed best suited to the French people on the basis of experience; they recognized that laws could not be inflexible "but must be adapted to the character, the habits, and the situation of the people for whom they are drafted."

Later changes and adaptations. No important changes were made in the Civil Code from 1804 to 1880, except the repeal of divorce in 1816, when a Catholic monarchy was restored. The political and legislative power was held by the bourgeoisie, and they were entirely satisfied with the basic principles of the code, which favoured individualism and free will. In fact, from 1804 until the enactment of the constitution of the Third Republic in 1875, the Civil Code remained the law of France despite several changes in political regimes. Jurisprudence was centred upon it; in both teaching and writing, scholars discussed it article by article. The courts fulfilled the role that the drafters had stressed for them; imbued with the spirit of the code, they applied its general rules to particular cases.

Changes during the Third Republic

The social atmosphere changed during the Third Republic when universal suffrage gave the labouring class an influence on legislation. Faith in liberalism was shaken, and the idea grew that the state should intervene to protect the weak. Statutes increased in number. This movement was accentuated by World Wars I and II, when a mass of emergency regulations had to be passed, and the power of the state to encroach on private interests for the sake of the community was increased.

Subsequent amendments to the code revealed two trends: first, greater individualism in family law; second, qualification of individual rights for the sake of social interests—what has been called "socialization" of the law.

Adaptation of the law to new social needs was not made by statute alone: the courts, to a certain extent, adjusted the law to modern circumstances. They did this, however, while maintaining a consciousness of their subordinate position. They recognized that, as a general rule, basic changes were the province of the legislature and not of the judge, though this did not prevent them from gradually adapting the law to the modern conditions of life.

Legal learning also had a role. A number of important statutes were drafted by commissions that included judges, professors, and lawyers; and authors often suggested to the courts new developments in the application of rules of law. Although most of the statutes passed during the 19th and 20th centuries were left outside the code, they continued to be published with the new editions of the code.

By the middle of the 20th century, it had become apparent that the code should be revised. This task was entrusted to a commission, which produced several important drafts. The effort to replace the old code with a completely new one was halted when Charles de Gaulle came to power in 1958. Revision has since occurred only on a sporadic, piecemeal basis, except for the sections concerning family law, which have been thoroughly reformulated. The continuing development of multinational institutions such as the European Communities may require readjustments on a broad scale. This seems particularly likely in the field of commercial law, where company law has become the first target for harmonization, and perhaps even in contract law.

The main categories of French private law. The French Civil Code uses many of the categories that were developed in ancient Rome, but its law is that of its own time.

Marriage viewed as basic to civilized society

Marriage and family. The drafters of the French Civil Code regarded marriage as the basic institution of a civilized society. Taking into account the variety of religious attitudes in France, they decided that only marriage ceremonies celebrated before secular officials should be legally valid. This did not deprive ministers of the various faiths of the right to celebrate religious marriage ceremonies, but these were devoid of any legal effect and had to take place after the secular ceremony in order to avoid any risk of confusion. Parental control over children's marriages was partially restored; consent was required for sons under 25 and daughters under 21. After 1900 the formalities of marriage were lessened and parental control over it curtailed. Twentieth-century statutes gradually reestablished the revolutionary rule that the consent of the parents was not necessary when the parties were over 21. In 1974 the age of majority for this and other purposes was reduced to 18.

In France under the ancien régime the family had been centred upon the husband, whose strong authority and powers were inherited from the Roman paterfamilias (head of family). Although the Revolution proclaimed women to be equal in rights with men, it did little to implement this view in law. The drafters of the code saw no reason to modify the traditional situation, and Napoleon himself favoured subordination of the wife to the husband. The code expressly stated that she owed him obedience. With very few exceptions, she had no legal capacity to act. Without the written consent of her husband, the wife could not sell, give, mortgage, buy, or even receive property through donation or succession. Statutes in the 20th century, however, severely diminished the authority of the husband over his wife and endowed her with full legal capacity. In 1970 the old language stating that "the husband is the head of the family" was abandoned in favour of a new principle of joint family decision-making power, which did not, however, extend to the management of community property.

Rights of husband and wife

In recent years matrimonial property regimes have been revised in numerous countries, the tendency being toward a partnership in property acquired after the marriage, with each party retaining control over the property he or she had before the marriage. Although the Napoleonic Code provided for a statutory regime (if no particular marriage contract had been made), under which all chattels and earnings of the spouses would be community property to be shared equally between them or their heirs at the dissolution of the marriage, the husband was vested with all active powers, even over his wife's property. In 1965 movables owned by either spouse before marriage were excluded from the community fund, which now, in the absence of an agreement to the contrary, consists only of the fruits of the spouses' work or frugality during marriage. With the acquisition of legal capacity in the early part of the century, a French wife was free to manage and dispose of her own earnings and property, but it was not until 1985 that the long predominance of the husband in the management of the couple's common property was replaced by a system of equal comanagement.

Divorce. Divorce was first introduced into France after the Revolution. It was made very easy and was even allowed by mutual agreement.

The drafters of the code decided that since many persons were not prevented by religious conviction from seeking divorce, it was not for the legislator to prevent unhappy spouses from terminating their marriages and from entering new legal unions. Divorce, therefore, was allowed, but only within strict limits, so that "the most sacred of contracts should not become the toy of caprice." The only grounds for divorce were adultery, sentences for the most serious crimes, excesses such as gambling habits and expenditures, cruel treatment, or serious insult. Mutual agreement was added under the personal pressure of Napoleon, already intent on divorcing his first wife, by whom he had no child. But the procedure of divorce by mutual agreement was extremely long, complicated, and costly, and no second marriage could take place within six years thereafter.

Divorce was repealed in 1816 after Napoleon's fall and the restoration of the monarchy, and it was not reintroduced until 1884. From 1884 to 1975, divorce was permitted only on the grounds of adultery, conviction of a serious crime, and cruelty. Divorce by mutual agreement was not reinstated until 1975, when a comprehensive reform of the divorce law permitted a marriage to be terminated by consent or by petition of one spouse unopposed by the other, or when the marriage had broken down after six years of separation or after six years of mental illness of one spouse, in addition to the traditional grounds of fault.

Succession and gifts. The Napoleonic Code adopted many of the ideas of the Revolution concerning succession. But its formulators tempered them with exceptions and combined them with ideas from the ancien régime.

The revolutionary law on intestate succession (succession without a valid will) relied upon two basic principles: (1) that no distinctions be made within the estate of the deceased, land and chattels being treated in the same way and no account being taken of the origin of landed property; and (2) that equal parts be given to all heirs of the same degree of kindred, the advantages accruing through some customs to the firstborn or to male children being abolished. Using these two principles, the code provided that an estate should devolve first of all upon the children and other descendants. If heirs of one degree died before others of the same degree and left children, representation (the principle that the children of a deceased heir inherit his share) applied. In other cases distribution was made per capita, with equal shares going to those heirs of equal degree. Illegitimate children could inherit from their parents, but they received less than legitimate children and could not cut out either the deceased's own parents or his brothers and sisters. Through reforms in the 1970s, the rights of illegitimate children to succeed to their parents have largely been assimilated to those of legitimate offspring.

Rights of heirs and surviving spouse

According to the code, the spouse could succeed only if there were no persons who were related to the deceased up to a degree specified by law. A surviving wife was, thus, in a poor position if no gift or legacy had been made to her, though under the statutory matrimonial regime she received half of the community property into which all chattels of both spouses fell. The rights of the surviving spouse have been increased at various times during the 20th century. Today the surviving spouse is entitled to at least the usufruct (similar to a life interest) of one-quarter of the property left by the deceased. The survivor inherits half of the estate if there are no children and if there are surviving ascendants on only one side of the deceased's family. If the decedent leaves no blood relatives within a certain degree of kindred, the surviving spouse receives the entire estate.

Wills may be formal or informal. Unwitnessed wills are valid, provided that they are written throughout, and dated and signed, by the testator's own hand. Wills are effective upon the death of the testator and do not need to be probated. Freedom to dispose of property by will or by gift is limited in order to protect children and other descendants as well as parents and grandparents, who have to be allowed a certain proportion.

Property. The intricate system of obligations and rights inherited by the ancien régime from feudalism was rejected by the Revolution, which restored a system patterned on that of Roman law.

Concept of movables and immovables

The only classification of goods is the basic one of immovables (which are defined as having a fixed place in space) and movables (which include all goods that are not immovables). In contrast to the "feudalist" complexities in common law, the normal relationship between persons and things is ownership, which is defined as a complete, absolute, free, and simple right. But, as in the law of other modern nations, the use of property is subject to many kinds of restrictions imposed in the public interest. Usufructs, or servitudes, are possible, but rights in an estate never require the person in whom they are vested to do anything. The code states that a servitude "is a charge laid on an estate for the use and utility of another estate belonging to another owner," and it emphasizes that "servitudes do not establish any pre-eminence of one estate over another." Title in land may be acquired within 10 or 20 years if the possessor believed, in good faith, that he was the real owner. Furthermore, the bona fide purchaser of movable property immediately becomes its owner, and nobody can prove a better title against him unless the property has been lost or stolen.

The section on mortgages in the Civil Code was weak. An excellent statute of the revolutionary period was developed in 1798 to set up a system of registration for all transfers of land titles and real estate mortgages. It enabled a buyer of land to ascertain whether he was buying from a regular owner and whether the land was mortgaged; if it was, the buyer could clear his title by offering the price to the mortgagee.

The drafters of the code maintained this system of compulsory registration, but only for gifts and for contractual mortgages. Sales of real estate and a number of legal mortgages were not subject to registration. This gap left prospective creditors or buyers with insufficient information. It was only after reforms were made in 1855, 1935, 1955, and 1967 that there was a comprehensive, but still not fully reliable, system of publicity for mortgages and conveyances of immovable property.

Contracts and torts. The French Revolution brought no changes into the law in this relatively nonpolitical field. The drafters of the code merely restated the law that had developed during the course of centuries and that authors already had analyzed.

Informality and freedom of contract

The basic principles of contract law are informality and freedom; the latter is limited, however, when demanded by public policy. The code states that "agreements legally entered into have the effect of laws on those who make them." The entire matter of torts is dealt with in only five short articles. The general basis for liability is the following: "Any act of a person that causes injury to another obligates the person through whose fault the injury occurred to give redress." The subsequent articles in the code regulate liability for damages caused by things, animals, children, and employees. It was left to the courts to work out a complete system of tort law based on these few articles.

THE GERMAN SYSTEM

Roman law, as embodied in the Corpus Juris Civilis, was "received" in Germany from the 15th century onward, and with this reception came a legal profession and a system of law developed by professionals (*Juristenrecht*). Roman law provided the theoretical basis for legal progress that culminated in the work of the scholars of the 19th century. Under this tradition, the legal process has been viewed in Germany as the application of more or less generally formulated rules to individual cases. German courts traditionally have not been as dominant in developing the law as have their counterparts in the common-law countries. Roman law provided tools to strengthen sovereignty, as well as the correlative ideas that the legislative function is a state monopoly and that the responsibility for the development of law rests with a legally trained, state-controlled bureaucracy rather than—as in 18th- and 19th-century England—with a combination of gentry and leaders of the bar. German judges traditionally have been university-trained experts under the authority of the state and the anonymity of the court. In the post-World War II period, however, West German judges have steadily assumed a more active role, especially in constitutional law.

Contrasts between the German and French codes

The German Civil Code. Because the German Civil Code of 1896 came almost 100 years later than the code of France, its draftsmen profited from the intensive efforts of German scholars who had systematized, clarified, and modernized the law during the 19th century. As a result, the German code is markedly different from its French predecessor: its arrangement is more orderly, its language more precise, and its use more exacting.

The appeal of the German code is from lawyers to lawyers; the matter-of-fact, neutral tone contrasts with the livelier mood in which the French Civil Code was written. It does not try to teach men in a broad sense, but it emphasizes ethical imperatives. Good faith and fair dealing are to be observed in all affairs. Breaches of good morals, abuses of rights, and underhanded legal transactions are deprived of legal effect. The code was meant to fit the society of the turn of the century, but through the use of general clauses that leave the elaboration of specific norms to the judges, it has demonstrated an adaptability to new economic, cultural, and sociopolitical postulates.

Modern law in West Germany assumes that the proper form of society is that of a social democracy, which not only confers individual rights but also holds the state responsible for social welfare and individuals responsible for behaving in a socially responsible way. The former concern of German law with abstract concepts has given place to a more pragmatic approach that attempts to apply the scale of values of a pluralistic society, containing strong elements of a welfare state, while emphasizing civil liberties.

The main categories of German private law. The German Civil Code begins with the proposition that at birth every person acquires the capacity to exercise rights and to fulfill duties. A minor's interests are protected by a representative who acts in his name, and although certain legal transactions may be entered into at age seven, full legal capacity is not acquired until age 18 (formerly 21). Every person possesses the right, protected by an action in court, to freedom from personal injury and from attacks on individual dignity.

Marriage and family. Since 1875 marriage has required civil celebration by a registrar, who cannot be a priest. Celebration in church may follow the civil ceremony. Marriage can be declared null and void on application by one of the spouses or by the public prosecutor on various grounds, such as lack of form or affinity, but the consequences of such nullity approximate those of divorce: the children are not necessarily illegitimate. Since 1976, the sole ground for divorce has been the breakdown of the marriage, which is presumed if the spouses have lived apart for a year and are in agreement on the divorce, or if the spouses have lived apart for three years.

Rights of husband and wife

The provisions of the German Civil Code concerning the rights of women in marriage were less restrictive than those of the French Civil Code. After World War II nearly all rules contravening the principle of equality of men and women were repealed. The ordinary statutory marital-property regime, with the husband administering and using the wife's estate, was replaced in 1957 by a system of separate management and equal sharing in the value of acquisitions made during the marriage. Upon the death of one spouse the surviving spouse is entitled to a generous share in the estate. Care for the person and property of the children belongs to both spouses.

Succession. In contrast to Anglo-American law, the assets of the decedent pass directly to the heirs, who are determined by the rules of intestacy or by testamentary disposition. As a general rule, the estate does not pass through a stage of administration by an administrator or executor. The heirs are liable for the debts of the decedent with their own property but by taking appropriate steps may limit their liability to the assets of the estate. A testator may appoint an executor to perform certain functions in the settlement. A will may be unwitnessed, but then it must be entirely in the testator's handwriting. Public wills are either made orally before a public official, who records them, or set down in a document that the testator hands to the official with a declaration that it is his last will. Descendants and other close relatives, including the surviving spouse, cannot be deprived of more than one-half of their intestate shares.

Property. Property is declared to entail obligations of the owner to the community. This is particularly important in terms of farmland, which can be pooled and redistributed to make better use of machinery and to increase production. Every creation, transfer, encumbrance, or cancellation of a right in immovable property requires, in addition to the agreement of the parties, registration with the district court. A person who, in good faith, acquires an interest in land from the person registered is protected. In order to obtain title to a chattel from a person who does not own it, the transferor must have had possession, the transferee must have been in good faith, and the owner must not have lost possession involuntarily. But neither in the case of land nor in that of chattels is it required that the transfer to the transferee be for value. Even if the transferee acquires a title, he may be required to surrender the asset or to pay its value if the acquisition appears to be a legally unjustified enrichment.

Contract and delict. Parties are free to regulate their relations by contract, within limits set by express statutory prohibitions and by good morals. Strict limits are set to eliminate fraudulent practices by one of the contracting parties. In the case of a valid contract, the parties must observe the requirements of good faith, with ordinary usage taken into consideration. The determination of "ordinary usage" is left to the courts. This has been particularly advantageous given the rapidly changing conditions of the 20th century.

Compensations for breach of contract and injury

Unless the promisor can prove that a breach of contract has been caused in a way entirely outside his sphere of risk, he is liable for damages. But if the promisee chooses to do so, he may have the promisor ordered to complete the contract as long as it is not shown that this is impossible. The principle that "anyone who through an act performed by another or in any other way acquires something at the expense of that other without legal justification is bound to return it to him" is stated in broad terms, but it is cautiously applied by the courts.

In terms of delict, the German Civil Code provides that any person who intentionally or negligently injures unlawfully the life, body, health, property, or any other absolute right of another person is bound to compensate him for any damage arising therefrom. Damages also are due for harm caused by the violation of a statute meant to protect others and for harm caused intentionally and immorally. If a public officer violates his statutory duty, court remedies are readily available against the government.

OTHER SIGNIFICANT CODIFICATIONS

Swiss and Italian codes

Swiss law. Shortly after German law was codified, Switzerland followed suit. Its Civil Code of 1907, together with a separate Code of Obligations, went into effect in 1912. These new federal codes superseded the earlier codes of the separate cantons (which had generally been patterned after the Austrian or the French model). The Swiss draftsmen took advantage of earlier experiences with codification technique—drawing especially upon both the Code Napoléon and the German code. The Swiss Civil

Code, which exists in German, French, and Italian versions of equal authority, is a masterly attempt at summarizing and systematizing civil law and has influenced codification in countries as diverse as Brazil and Turkey.

Italian law. The French code was introduced into parts of Italy during the Napoleonic conquests. Even after the collapse of Napoleon's empire, when French law was abrogated, the Code Napoléon still served as the model for the new codes of several Italian states. The new Civil Code for the Kingdom of Italy was enacted in 1865 while the peninsula was being united politically. Its structure and content were reproductions of the French Civil Code. Unlike France and Germany, which have occasionally tried to draft new codes but still have not replaced their original ones, Italy succeeded in introducing a reformed code in 1942, during the Fascist era. This code remains in force, with its only amendments due mainly to changes in political regimes. In comparison with the Civil Code of 1865, Italy's code of 1942 appears inspired by less individualist views—for example, in property law and in labour legislation the social aspects of the law are stressed.

The Japanese code

Japanese law. After the Meiji Restoration of 1868, which abolished feudal privileges and restored titular power to the emperor, the leaders of the new government sought to construct an economic, political, and legal structure capable of commanding respect internationally. The introduction of Western law was one element of a wholesale importation of things Western. In legal matters, the Japanese took for models the systems of continental Europe, especially the German. The drafters of the Japanese Civil Code of 1898 surveyed many legal systems, including the French, Swiss, and common-law, and they took something from each. Their final product was, however, best characterized as following the first draft of the German Civil Code. In subsequent developments, the Japanese legal system remained true to these sources. In 1947 revisions of code provisions dealing with family law and succession, which had reflected traditional Japanese attitudes, completed the transition of Japanese civil law to the continental European family of laws.

In several respects, however, Japanese law is closer to that of the United States than to European models, especially in matters of public and constitutional law. This is largely a result of the post-World War II occupation and of subsequent contacts with U.S. legal thinking and education. From the perspective of the rules and institutions of private law, the Japanese legal system remains closer to the civil law of Europe than to the common law of the United States. In many ways, nevertheless, the Japanese legal order differs markedly from all Western legal orders.

The fact that Japanese law is not the product of organic evolution suggests that the role of law in modern Japanese society differs markedly from its role in Western societies. In Japan, law plays a far less pervasive role in the resolution of disputes and in the creation and adjustment of rules regulating conduct. The size of the Japanese bar is small, and extralegal methods of resolving disputes continue in large measure.

For many purposes the Japanese family transcends husband, wife, and dependent children. The notion that a business is analogous to a family unit also persists and colours all labour relations, especially in small and middle-sized firms. In the relatively homogeneous Japanese society, social status carries heavy obligations, and community pressure is extremely powerful.

Thus, although Japan early adopted a version of the German Civil Code, it did not adopt the Germans' strong consciousness of legal rights. In many areas of Japanese life it is still difficult to predict whether a dispute will be settled under legal standards, and it is often impossible to know whether a person will enforce those rights that are legally available to him. The concepts pervasive in Western law—that the legal consequences of a particular conduct should be predictable before the conduct has occurred, that in any dispute the courts should give full effect to claims (a plaintiff receiving all or nothing), and that individual disputes should be resolved without considering the parties' social and economic background—have not penetrated deeply into Japanese law. In contrast,

Alternatives to legal action

facilities for conciliation are used to promote adjustment in terms of nonlegal considerations: local police stations provide conciliation rooms; elders act as go-betweens. Compromise based on legally irrelevant considerations is encouraged, and disputes are often resolved by techniques that fall outside formal law.

Despite deliberate governmental efforts to limit the number of legal professionals, the continuing westernization of Japan may force Japanese law to play a role fully comparable to the role of law in the West; the sociological supports essential to the continued vitality of the traditional Japanese conception of law are clearly being undercut by Japan's shift to a highly urban, mechanized society.

(M.Rh./M.A.Gl.)

Common-law systems

THE HISTORICAL RISE OF COMMON LAW

English common law, or the body of customary law embodied in reports of decided cases, originated in the early Middle Ages in the decisions of local courts, which applied what Blackstone called "the custom of the realm from time immemorial" and practical reason to everyday disputes with the aid of but few formal enactments. Until the late 19th century, English common law continued to be developed primarily by judges rather than legislators.

The common law of England is in fact largely a Norman creation. The Anglo-Saxons, especially after the accession of Alfred the Great (871), developed a body of rules resembling those being used by the Teutonic peoples of northern Europe. Local customs governed most matters, while the church played a large part in government. The concept of crimes originated in this era, but they were treated as wrongs for which compensation was made to the victim.

The Norman Conquest of 1066 brought a practical end to the Saxon laws, except for some local customs. All of the land was allocated to Norman feudal vassals of the king. Serious wrongs were regarded mainly as public crimes rather than as personal matters, and the perpetrators were punished by death and forfeitures of property. Government was centralized, a bureaucracy built up, and written records maintained. Royal officials roamed the country, inquiring into the administration of justice. Church and state were separate and had their own law and court systems. This led to centuries of rivalry over jurisdiction, especially since appeals from church courts, before the Reformation, could be taken to Rome. Some elements of Saxon practice lingered, including trial by ordeal (by burning the hand, for example), which was retained until 1215. Outlawry, a Saxon procedure whereby a fugitive was placed outside the protection of the law, was retained for centuries to deal with people who fled from justice. Gradually, however, new procedures took the place of these crude devices.

Remnants of Saxon law

The Normans spoke French and had developed a customary law in Normandy. They had no professional lawyers or judges; instead, they used "clerks," or literate clergymen, to act as administrators. Some of the clergy were familiar with Roman law and the canon law of the Christian Church. Canon law was adopted by the English church, but the Normans resisted any attempt to introduce Roman law, which was applied only to certain claims under wills in the church courts, to marine disputes in the admiralty courts from the 14th century, and to military law. Norman custom was not simply transplanted to England, and a new body of rules, based on local conditions, grew up.

The feudal land law. At the critical formative period of common law, the English economy depended largely on agriculture. Wages and profits were important only in commercial centres such as London, Norwich, and Bristol. Political power was rural and based on landownership. Landowners voted at elections as Parliament evolved, and they acted as sheriffs and magistrates and sat on juries.

Land was held under a chain of feudal relations. Under the king came the aristocratic "tenants in chief," then strata of "mesne," or intermediate tenants, and finally the tenant "in demesne," who actually managed the property. Each piece of land was held under a particular condi-

Conditions of tenure

tion of tenure; that is, in return for a certain service or payment. An armed knight, for example, might have to be provided to serve in the king's armies for a certain period each year. Nonmilitary service, such as making deliveries of grain, was often substituted for the uncertain obligations of knight service. Periodic services tended to be commuted into fixed annual payments, which ceased, under the impact of inflation, to have much value. The "incidents," or contingency rights, however, such as the right of the feudal lord to take the land if the tenant died without heirs and his rights regarding wardship and marriage of the tenant's infant heirs (that is, his rights to compensation for exercising wardship or granting permission to marry) were assessed at current land values and remained important.

Succession to tenancies was regulated by a system of different "estates," or rights in land, which determined the duration of the tenant's interest. Land held in "fee simple" meant that any heirs could inherit (that is, succeed to the tenancy), whereas land held in "fee tail" could pass only to direct descendants. Life estates (tenancies lasting only for one person's lifetime) could also be created. Title to land was transferred by a formal ritual rather than by deed because the population was largely illiterate. Few elaborate rules about the terms by which land was held could be agreed upon in such circumstances, so statutes were passed to regulate matters of detail. The life tenant, for example, was forbidden in the 13th century to use the property in such a way as to damage it or to cause it to deteriorate unless the grant specifically allowed it, and the tenant "in tail" was forbidden to ignore the system of descent laid down for his property. The common-law judges devoted themselves to working out the proper rules to apply to all of these estates and tenures.

Primogeniture—*i.e.*, the right of succession of the eldest son—became characteristic of the common law. It was designed only for knight-service tenures but was inappropriately extended to all land. This contrasted with the widespread practice on the Continent, whereby all children inherited equal shares.

Development of a centralized judiciary. The unity and consistency of the common law were promoted by the early dominant position acquired by the royal courts. A single royal court, the King's Court (Curia Regis), was set up for most of the country at Westminster, near London. Whereas the earlier Saxon *witan*, or king's council, dealt only with great affairs of state, the new Norman court assumed wide judicial powers. Its judges (clergy and statesmen) "declared" the common law.

By straining the interpretation of a statute, royal judges greatly reduced the jurisdictions of local courts. With their increased powers, royal judges went out to provincial towns "on circuit" and took the law of Westminster everywhere with them, both in civil and in criminal cases. Local customs received lip service, but the royal courts controlled them and often rejected them as unreasonable or unproved: common law was presumed to apply everywhere until a local custom could be proved. This situation contrasted strikingly with that in France, where a monarch ruled a number of duchies and counties, each with its own customary law, or of the situation in lands such as Germany and Italy, which were divided into independent kingdoms and principalities with their own laws.

Dissemination of common law by itinerant judges

This early centralization also removed the need for England to import a single advanced foreign system of law, a need that led to the reception of Roman law in Europe after the decline of feudalism. The expression common law, devised to distinguish the general law from local or group customs and privileges, came to suggest to citizens a universal law, founded on reason and superior in type.

In the 13th century the common central court split into three courts—Exchequer, Common Pleas, and King's Bench. Although the same law was applied in each, they vied in offering better remedies to litigants in order to increase their fees.

The court machinery for civil cases was built around the writ system. Each writ was a written order in the king's name issued at the instance of the complainant and ordering the defendant to appear in the King's Court or ordering some inferior court to see justice done. It was based on a form of action (*i.e.*, on a particular type of complaint, such as trespass), and the right writ had to be selected to suit that form. Royal writs had to be used for all actions concerning title to land.

Bracton and the influence of Roman law. Under Henry III, who reigned from 1216 to 1272, an assize judge (*i.e.*, an itinerant judge of the periodical local assize courts), Henry de Bracton (originally Bratton), prepared an ambitious treatise known as "Bracton." It was modeled on the order of the 6th-century Roman legal classic, the Institutes of Justinian, and shows some knowledge of Roman law. Its English character derived from the space it devoted to actions and procedure, to the reliance on judicial decisions as declaring the law, and to the statements limiting absolute royal power. Bracton abstracted several thousand cases from court records (plea rolls) as the raw material for his book. The plea rolls formed an almost unbroken series from 1189 and included the writ, pleadings, verdict, and judgment of each civil action.

Early statute law. Edward I has been called the English Justinian because his enactments had such an important influence on the law of the Middle Ages. Edward's civil legislation, which amended the unwritten common law, remained for centuries as the basic statute law. It was supplemented by masses of specialized statutes that were passed to meet temporary problems.

Four of Edward's statutes deserve particular mention. The first Statute of Westminster (1275) made jury trial compulsory in criminal cases and altered land law. The Statute of Gloucester (1278) limited the jurisdiction of local courts and extended the scope of actions for damages. The second Statute of Westminster (1285), a very long enactment, confirmed the estate tail in land, which had often been linked with the maintenance of titles of honour; made land an asset for purposes of paying judgment debts (debts judged to exist by a court); liberalized appeals to high circuit courts; improved the law of administration of assets on death; and created a new form of action, action on the case, that gave broad approval to the creation of new remedies for new types of contract and tort cases. The Statute of 1290 (Quia Emptores) barred the granting of new feudal rights, except by the crown, and made all land held in fee simple freely transferable by denying interference by relatives or feudal lords.

Major statutory enactments of Edward I

In modern times the statutes issued prior to 1285 are sometimes treated as common law rather than statute law. This is because these laws tended to restate existing law or give it a more detailed expression. They explained what the law was, but they did not make an entirely new law; some authorities, in fact, doubted whether governments had the right to change ancient customs at all. In addition, judges did not always adhere closely to the words of the statute but tried to interpret it as part of the general law on the subject. Prior to the rise of the House of Commons in the 13th century, it also was difficult to distinguish acts of Parliament from the less binding decisions or resolutions of the royal council, the executive authority. Some statutes were passed but never were put into force, while others seem to have been quietly ignored.

The second Statute of Westminster, however, clearly made new law and allowed time for citizens to study its provisions before it came into force. Even so, this statute was freely interpreted by the courts, who read into it things that were not in the text.

Growth of Chancery and equity. Since legal rules cannot be formulated to deal adequately with every possible contingency, their mechanical application can sometimes result in injustice. In order to remedy such injustices, the law of equity (or, earlier, of "conscience") was developed. The principle of equity was as old as the strict common law, but it was hardly needed until the 14th century, since the law was still relatively fluid and informal. As the law became firmly established, however, its strict rules of proof began to cause hardship. Visible factors of proof, such as the open possession of land and the use of wax seals on documents, were stressed, and secret trusts and informal contracts were not recognized.

Power to grant relief in situations involving potential injustices lay with the king and was first exercised by the entire royal council. Within the council, the lord chancellor, a leading bishop, led the meetings and, by 1474, dealt personally with petitions for relief. Eventually the chancellor's jurisdiction developed into the Court of Chancery, whose function was to administer equity. Much of the work concerned procedural delays and irregularities in local courts, but gradually the power to modify the operation of the rules of common law was asserted.

Early grounds for equitable relief

The chancellor decided each case on its merits and had the right to grant or refuse relief without giving reasons. Common grounds for relief, however, came to be recognized. They included fraud, breach of confidence, attempts to obtain payment twice, and unjust retention of property.

Proceedings began with bills being presented by the plaintiff in the vernacular language, not Latin; the defendant was then summoned by a writ of subpoena to appear for personal questioning by the chancellor or one of his subordinates. Refusal to appear or to satisfy a decree was punished by imprisonment. Because the defendant could file an answer, a system of written pleadings developed.

Inns of Court and the Year Books. During Edward I's reign (1272–1307) the office of judge was transformed from a clerical position into a full-time career. Admission to the bar (*i.e.*, the right to practice as a barrister before a court) was made conditional on the legal knowledge of the applicant so that law began to emerge as a profession which required permanent institutions and some kind of organized education.

As the legal profession grew, the more experienced barristers were admitted to the dignity of serjeant-at-law and later banded together with the judges, who were appointed from their ranks, at Serjeants' Inns, in London. There, burning legal problems were informally discussed, and guidance was given to all concerning the decisions of actual or likely cases. The four Inns of Court (Gray's Inn, Lincoln's Inn, Inner Temple, and Middle Temple) evolved from the residential halls of junior barristers to become the bodies officially recognized as having the right to admit persons to the bar. Education consisted of attending court, participating in simulated legal disputes (moots), and attending lectures (readings) given by senior lawyers.

Bracton's work was adapted for purposes of study for a time, but it soon became outdated. Bar students therefore had to make notes in court of actual legal arguments in order to keep abreast of current law practices. These notes varied widely in quality, depending on the ability of the notetaker and the regularity of his attendance, and from about 1290 they seem to have been copied and circulated. In the 16th century they began to be printed and arranged by regnal year, coming to be referred to as the Year Books.

The Year Book reports were written in highly abbreviated law French. They did not always distinguish between the judges and barristers and often simply referred to them by name. The actual judgment also was often omitted. Previous decisions were not generally binding, but great attention was paid to them, and it appears that the judges and barristers referred to earlier Year Books in preparing their cases. Thus, case law became the typical form of English common law.

Effects of feudal conflict on the law

The dynastic Wars of the Roses in the latter part of the 15th century led to a practical breakdown of the legal order. Powerful hereditary aristocrats in the country, backed by private armies, and dominant commercial families in the towns were beyond the effective reach of the royal writ. When legal proceedings were possible, they were often manipulated or frustrated by the crown's "overmighty subjects," who intimidated and corrupted justices, sheriffs, juries, and witnesses.

Thus the years preceding the Tudor period were a time of insecurity and stagnation, a "Gothic age" in which lawyers tried to consolidate the law but made no new advances. Parliamentary authority also was weakened, and the royal council was called on more and more to rule the country and try to maintain order.

The rise of the prerogative courts. The accession of Henry VII in 1485 was followed by the creation of a number of courts that stood outside the common-law system that Henry II and his successors had instituted. These newer courts were described as prerogative courts because they were identified with the royal executive power, although some of them had a statutory origin. Thus, the Council of the North at York was set up by statute in 1537, and the Council of Wales and the Marches at Ludlow was confirmed by statute in 1543, though both had been preceded by older prerogative courts in those "frontier" regions. The Court of Requests (see below) was given regular status by an administrative action in 1493. The Court of Star Chamber, once thought to have been given its authority by a statute of 1487, is now believed to have evolved from the royal council, which began acting as a judicial committee in the early 16th century. All these courts rested on the comparative authority and efficiency of the council in times when regular courts were unable to operate properly.

The Court of Requests

In the Court of Requests, which had counterparts in France, the costs of procedure were lower than in common-law proceedings; it was designed to accommodate small civil claims by the poor. The judges of the court were styled masters of requests, and they had many other duties, which often caused delays. The court flourished in the 17th century until the Civil War (1642–51), when the procedure by which it operated was abolished. Its example of offering a simple, cheap procedure was imitated by several statutory courts that were set up in towns in later times, also known as courts of requests.

Whereas the common-law courts punished "hanging crimes," such as murder and robbery, the Star Chamber dealt with more sophisticated offenses, such as forgery, perjury, and conspiracy. Fines and sentences of imprisonment were the usual punishment. Common-law judges, lay peers, and bishops sat on this court, which also exercised civil jurisdiction. It lost its original popularity when the early Stuart kings used it to stifle political opposition, and its name eventually became synonymous with repression. It was abolished in 1641, and most of its jurisdiction was absorbed by the common-law courts in 1660.

The rather specialized High Court of Admiralty developed under royal prerogative in the 14th century; a statute of 1391 prohibited it from meddling in cases not arising at sea. In Tudor and early Stuart times, however, it exercised a wide commercial jurisdiction. After the Civil War it was confined exclusively to trying maritime disputes.

Further Roman-law influences. As described above, the common law had begun to break down in the 15th century. Abroad, law was in a state of flux. The customs of northern France were codified in 1453, and modified Roman law became a main source of imperial German law in 1495 and of Scots law in 1532. At the same time, the scope of canon and Roman law in England was increasing. Admiralty law, for example, drew on Greek, Roman, and Italian law and used documents drawn up in continental form, and the crimes of forgery and libel tried in Star Chamber were based on Roman models. Ecclesiastical courts applied canon-law rules based on Roman law, for example, to wills and marriages. The Councils of Wales and the North also used Roman law. All of these bodies competed with common-law courts for jurisdiction over the same cases and followed a written procedure modeled after that still being used on the Continent. Roman law and canon law, furthermore, were taught at Oxford and Cambridge, which gave doctorates to the practitioners in these courts.

One of the accusations reportedly made against Thomas Cardinal Wolsey, who fell from favour in 1529, was that he planned to introduce Roman law into England; Wolsey did appoint many clergy to the Council of the North and as justices of the peace. The 19th-century English legal historian F.W. Maitland discussed this legal crisis in a famous essay on English law and the Renaissance. Maitland ascribed the survival of the common law, in part, to the solid front presented by the Inns of Court, which trained lawyers practically and not theoretically. The English law tradition did not depend on abstract scholarly commentaries but on detailed judicial rulings about specific points of law arising in practice.

The influence of Roman-law ideas, however, was probably greater than generally admitted. The actions of trespass and disseisin (dispossession) had Roman analogies, and the estate tail was clearly influenced by a law made by Justinian. The equitable remedy of injunction had analogies in canon law, and the law of redemption of mortgages may have been related to the usury laws, which forbade making excessive profits from loans. The law of trusts and deceit resembled the breach of faith of the church courts. Continental mercantile law, which contained Roman-law elements, was absorbed into English law as it stood. Continental law also contributed to some of the rules of contract, such as the effect of mistake, and the Roman concept of fault played a part in the law of negligence. Many old European legal ideas, in fact, survived longer in England, where they escaped being eliminated in codifications, than in Europe.

Coke as a champion of the common law

An account of the development of common law in the Tudor-Stuart period would be incomplete without mention of Sir Edward Coke. Coke, who combined a distinguished career as a barrister and a judge, produced a wealth of legal writings. In 1606 he risked removal from the office of chief justice by challenging the exaggerated claims of the royalist party to prerogative powers outside of the common law. He disapproved of legislation by proclamation, of dispensation from the law in individual cases, and of the mushrooming jurisdictions of the prerogative courts. He helped draft the Petition of Right in 1628.

Coke's 11 volumes of *Reports* appeared between 1600 and 1615, and two posthumous volumes followed. Coke commented, rather than reported, but he was careful to supply a copy of the court record of each case. As the only formal series of collected law cases available at the time, his reports formed the main source for the citation of cases for many years. His four volumes of *Institutes of the Lawes of England,* published between 1628 and 1644, dealt with the law of real property (*Coke on Littleton*), the medieval statutes, the criminal law (pleas of the crown), and the jurisdiction of the courts.

Coke was no objective historian but an open advocate of the common law. Though he was old-fashioned and at times in error, his greatest works restated the common law in acceptable form and did much to save it.

Further growth of statute law. The Tudors made use of proclamations by the king to invoke emergency measures, to establish detailed regulations, especially on economic matters, and to grant royal charters to trading companies. Parliament passed laws of a political character, such as those enforcing the king's supremacy over the new established church. Statutes also regulated imports and exports, farming, and unfair competition. A law of 1562–63 regulated apprenticeships and provided for annual wage fixing by magistrates in accordance with the cost of living.

Among other statutory innovations were the Statute of Monopolies of 1623, which confirmed that monopolies were contrary to common law but which made exceptions for patentable inventions; a statute of 1601 that became the basis of the privileges enjoyed by charitable trusts; and the series of Poor Laws, which were enacted in the late 16th century to remedy the neglect of the poor caused by the dissolution of the monasteries.

In 1540 legal actions to recover land were subjected to time limits, and in 1623–24 the principle of limitation of actions by lapse of time was introduced into the law of contract and tort.

During the Commonwealth (1649–60) many reform projects were drafted; although they anticipated 19th-century reforms, none of them was carried out. These reforms included supplying counsel to prisoners, modernizing the land and law procedure, and permitting civil marriages.

The outstanding enactment of the later Stuart period was the Statute of Frauds of 1677. As a response to the growth of literacy and the prevalence of perjury and fraud, wills and contracts for sale of land or goods (of more than a certain amount) were required to be in writing. Though drafted by eminent judges, the statute was to require endless interpretation.

Further development of equity. Although one eminent contemporary observer, the legal historian John Selden, regarded the fate of a lawsuit in Chancery as varying with the chancellor's personality, the types of suits that would be granted relief had eventually become fairly clear. Precedent was being followed, and law reports of equity decisions and books on equity began to be published.

In 1615 the King declared that the Chancery was to retain its traditional superiority over the common-law courts but only in areas in which its authority was well recognized. If the applicability of equity was in doubt, the common law was to be followed.

Developments in the law of trusts

The main development in this period was in the law of trusts (see PROPERTY LAW). In medieval England, from the 14th century, most land was held "to uses"; *i.e.,* by nominees for the true owners. This situation may have been partly due to devices used to evade taxation, but it also enabled wills of land to be made. "Death duties" were payable if a man died while he was the legal proprietor; by transferring the land to another person, these could be avoided. Wills of land were not allowed before 1540, but the use of land could be transferred to another person while the owner was still alive, as long as the transferee observed the owner's wishes regarding the land while the owner lived. The beneficiary of such a trust usually stayed on the land as apparent owner, though the trustee held the legal title. A statute of Richard III, however, allowed the beneficiary to transfer the property, and in 1535–36 the Statute of Uses eliminated the middleman and revested the legal title in the beneficiary. The device of the use was exploited to create new and complicated legal interests in land. The old use was revived as the modern trust in Chancery, first for trusts involving money and leases and finally for trusts of land itself. The spur was the desire to separate the legal and beneficial titles, especially when the beneficiary was young or inexperienced. But the trust was adapted to many other ends, such as giving property to clubs and other unincorporated bodies and to churches.

THE MODERNIZATION OF COMMON LAW IN GREAT BRITAIN

Influence of Blackstone. Of extraordinary influence in the development of common law and in its dissemination to other parts of the world was the most famous of English jurists, Sir William Blackstone. Born in 1723, he entered the bar in 1746 and in 1758 became the first person to lecture on English law at an English university.

His most influential work, the *Commentaries on the Laws of England,* was published between 1765 and 1769 and consisted of four books: "Persons" dealt with family and public law; "Things" gave a brilliant outline of real-property law; "Private Wrongs" covered civil liability, courts, and procedure; and "Public Wrongs" was an excellent study of criminal law.

Blackstone was far from being a scientific jurist and was criticized for his superficiality and lack of historical sense. The shortcomings of the *Commentaries* in these respects, however, were offset by its style and intelligibility, and lawyers and laymen alike came to regard it as an authoritative exposition of the law. In the following century the fame of Blackstone was even greater in the United States than in his native land. After the Declaration of Independence the *Commentaries* became the chief source of knowledge of English law in the New World.

Reform in the 19th and 20th centuries. *Bentham.* Following the social turmoil of the French Revolution and the economic upheaval of the Industrial Revolution, there were many demands for reforms to modernize the law. The most significant figure in the reform movement was the English Utilitarian philosopher Jeremy Bentham, who was prepared to reform the whole law along radical lines. A brilliant student, Bentham disliked the picture of the law that was presented in Blackstone's lectures. In 1769 he entered the bar, but since he was living on an inheritance, he never found it necessary to enter practice. He worked to make law less technical and more accessible to the people, but he was slow to complete or publish his writings, and not until 1789 did his basic work, *An Introduction to the Principles of Morals and Legislation,* appear.

Bentham attacked legal fictions and other historical anomalies. He advocated two basic changes in the legal system: in order to achieve the greatest happiness for the greatest number, legislators, rather than courts, should make the law; and the aims of law should vary with time and place.

The fame of the *Principles* spread widely and rapidly. Bentham was made a French citizen in 1792, and his advice was respectfully received in most of the states of Europe and America. Although he wanted most of all to be allowed to draw up a legal code for his own or some foreign country, his practical influence was far more indirect and derived largely from the diffusion of Utilitarian ideas during the 19th century.

Changes in procedure and criminal law. In England the restrictive framework of the separate forms of action in civil cases was replaced in 1852 by a new system of uniform writs of summons, and liberal amendment of pleadings was permitted. Fixed dates were established for trials. In 1933 jury trial was ended in civil cases, except in libel and a few other actions. Evidence acts of 1938, 1968, and 1972 simplified civil proof. A major trend in criminal procedure since the early 19th century has been better protection of the rights of the accused. Since 1836 the accused has been entitled to counsel and since 1898 has been allowed to testify on his own behalf. In 1903 provision for the state to pay for defense was made and since expanded, and in 1907 the right of appeal against criminal convictions was created. In 1967 verdicts by a majority of the jury were made possible, and restrictions were imposed on press coverage of preliminary hearings.

Increased protection of the rights of the accused

The 19th century saw the enactment of a series of statutes that codified the part of criminal law dealing with individual crimes, apart from homicide. Basic ideas have changed little, other than the fact that some modern statutes have imposed responsibility without fault and that corporations can now be held responsible for the acts of their management.

The rules of legal insanity were laid down in the 19th century and supplemented in 1957 by the limited defense of "diminished responsibility." Capital punishment was gradually ended for most felonies and was finally eliminated for murder by the Homicide Acts of 1957–65. In 1968 a new Theft Act replaced the rather crude medieval idea of larceny by a broader concept that resembles the Roman delict (offense) of theft. Experimentation has led to new remedies, one of these being the suspended sentence, which only has to be served if a further crime is committed.

Reorganization of the judiciary. The lay jurisdiction of the church courts ended in 1857, when the divorce and probate courts were set up. These merged into the High Court of Justice in 1875 as a result of the Judicature Acts of 1873–75, which reformed the civil courts. The Judicature Acts were much more than a regrouping and renaming of courts; they attempted to fuse law and equity by making available legal and equitable remedies in all divisions of the High Court and by providing that the equitable rule should prevail when conflicts arose. Common law and equity nevertheless preserved their separate identities, partly because of the different subject matter with which they often dealt and partly because lawyers persisted in maintaining the distinction.

In the late 19th century the three central courts of common law were amalgamated as the Queen's Bench Division, which to this day continues to try suits for damages. Since 1875 cases have been tried by a single judge (before 1933 with a jury), not by a full bench of judges.

After it became a division of the High Court in 1875, the Chancery not only dealt with equity suits but also administered the voluminous legislation on property, bankruptcy, succession, copyrights, patents, and taxation. Contested probate cases were transferred to the Chancery by the Courts Act of 1971.

Before the Courts Act criminal cases were tried two or three times a year at assizes or four times a year at quarter sessions in the provinces. As of January 1972 a system of provincial crown courts replaced these. Civil assizes were replaced by allowing the High Court to sit at certain cities. Small civil cases, tried at statutory county courts since 1846, are now regulated by an act introduced in 1984.

A remarkable feature of English criminal justice, as compared with most European systems, has been the continuing role of lay justices of the peace, who remain important despite the appointment of paid, legally trained magistrates in London and some of the larger cities, of barristers as recorders at borough quarter sessions, and of legally qualified chairmen at county quarter sessions. An important aspect of the magistrates' work has been their jurisdiction over young offenders, for whom special juvenile courts were first set up in 1908. The report of a royal commission on justices of the peace in 1948 strongly defended the position of lay justice against public criticism; its cautious recommendations as to the appointment of justices and as to the organization of their courts were largely put into effect by the Justices of the Peace Act (1949) and the Magistrates' Courts Act (1980). The Criminal Justice Administration Act (1962) extended the power of justices of the peace to try indictable offenses summarily. A series of statutes in 1972, 1973, 1977, 1981, and 1982 rendered the procedure more flexible, made detailed provision for penalties and their execution, and added a number of new offenses. In 1964 elementary judicial training for lay justices was introduced. These developments since 1948 show both the persistence in English law of ancient institutions and a preference for reforming rather than totally abolishing them.

Routes of appeal

A modern appellate court for civil cases in the High Court was set up in 1830 but was replaced in 1875 by a Court of Appeal consisting of special appellate judges. In 1907 a Court of Criminal Appeal was established, but it was merged into the Court of Appeal in 1966. A divisional court hears appeals from magistrates on points of law. A final appeal, subject to conditions, can be made to the House of Lords from all lower courts.

Reform in private law. Property law has been changed often. Wills are regulated mainly by a statute of 1837 (amended in 1982), and the freedom to disinherit has been curtailed by family provision acts of 1938, 1952, 1966, and 1975. Title to land is subject to a system of registration that has been gradually introduced under an act of 1925. Succession on intestacy (*i.e.,* in the absence of a valid will) for all kinds of property was unified in the same year. The law of leases has been modified by social legislation such as the numerous Rent (control) Acts, which protect residential tenants. The terms of trusts can be modified by the Chancery (since 1958), and a wider range of trustee investments has been allowed since 1961.

Grounds for divorce have been enlarged by a number of 20th-century statutes, culminating in the broad "breakdown of marriage" approach of the Divorce Reform Act of 1969, now the Matrimonial Causes Act of 1973 (as amended in 1984). Under this legislation a marriage may be terminated not only on traditional fault grounds but also when the parties have lived apart for at least two years and consent to divorce or when the parties have been separated for at least five years.

After several piecemeal laws addressed trade (labour) unions, a more comprehensive, but controversial, Industrial Relations Act was passed in 1971, requiring registration of unions and arbitration of disputes. This statute was repealed in 1974, but aspects of it were revived with considerable modification in 1980 and 1982.

In the field of tort, manufacturers' liability to consumers was established by case law in 1932 and later strengthened by legislation. Liability in libel has been cut down by many statutes. A law of 1945 introduced the Roman principle of apportioning damages when both parties are at fault.

Commercial law, with the Bills of Exchange Act (1882), Sale of Goods Act (1893 and 1979), the Unfair Contract Terms Act (1977), and consumer protection statutes in 1965 and 1974, has become primarily the domain of legislation. Arbitration, too, is regulated by statute.

THE DEVELOPMENT OF COMMON LAW IN THE UNITED STATES

The first English settlers on the Atlantic Seaboard of North America brought with them only elementary notions of law. Colonial charters conferred on them the traditional legal privileges of Englishmen, such as habeas corpus and the right to trial before a jury of one's peers, but there were few judges, lawyers, or lawbooks, and English court decisions were slow to reach them. Each colony passed its own statutes, and governors or legislative bodies acted as courts. Civil and criminal cases were tried in the same courts, and lay juries enjoyed wide powers. English laws passed after the date of settlement did not automatically apply in the colonies, and even presettlement legislation was liable to adaptation. English cases were not binding precedents. Several of the American colonies introduced substantial legal codes, such as those of Massachusetts in 1648 and of Pennsylvania in 1682.

The reception of English law

By the late 17th century, lawyers were practicing in the colonies, using English lawbooks and following English procedures and forms of action. In 1701 Rhode Island legislated to receive English law in full, subject to local legislation, and the same happened in the Carolinas in 1712 and 1715. Other colonies, in practice, also applied the common law with local variations.

Many legal battles in the period leading up to the War of Independence were fought on common-law principles, and half of the signatories of the Declaration of Independence were lawyers. The U.S. Constitution itself uses traditional English legal terms.

After 1776 anti-British feeling led some Americans to advocate a fresh legal system, but European laws were diverse, couched in foreign languages having unfamiliar turns of thought, and unavailable in textbook form. Blackstone's *Commentaries,* reprinted in America in 1771, was widely used, even though new English statutes and decisions were officially ignored.

In the 1830s two great judges, James Kent of New York and Joseph Story of Massachusetts, produced important commentaries on common law and equity, emphasizing the need for legal certainty and for security of title to property. These works followed the common-law tradition, which has never been fundamentally altered in the United States, except in Louisiana, where French civil law has survived.

American innovations. The American states saw law as a cementing force, and they used it to facilitate cooperation in the face of the hazards of nature and other difficulties arising in the development of the new continent. Special laws were developed to deal with timber, water, and mineral rights. Simple procedures were followed. Dogma was rejected in favour of personal experience and experiment, and old decisions soon became outdated. The pioneer spirit favoured freedom and initiative and distrusted central authority and a paternal government. Homespun local justice was preferred, as was the common sense of the local jury. For a time some of the colonies even tried to base their law on the Bible. But, even when English law reasserted itself, many of its institutions were rejected. On death intestate, for example, all of the children inherited land in America and not just the eldest son, as in England. Freehold title was the rule, not long leases under landlords. Church courts did not exist.

Growth of statute law and codes. After the War of Independence a drive to replace judge-made law by popular legislation was revived. In 1811 Jeremy Bentham proposed a national civil code to Pres. James Madison, but his proposal was premature. In the mid-19th century, the legal reformer David Dudley Field presided over the drafting of several codes and campaigned vigorously for the systematic, rational codification of U.S. law. Except for a code of civil procedure, which was widely copied, Field's codes found little acceptance in state legislatures. Field's civil code was adopted by five states, including California and New York, but the common-law tradition was so strong in these jurisdictions that the civil code became just another statute; it was read against the background of, and supplemented by, existing case law, rather than seen as a complete set of authoritative starting points for legal reasoning as were the continental civil codes. Louisiana, whose legal system is a hybrid of civil- and common-law elements, is the only American state that has a code in the civil-law sense. Despite the failure of the codification movement, U.S. law became increasingly statutory, so that by the late 20th century legislation predominated over judge-made law.

Codes of David Dudley Field

U.S. statutes are not construed so narrowly as those in England, and there is less reluctance to change the older law. Statutes are also regularly revised; for example, New York state has had a Law Revision Commission since 1934.

Equity and probate. In 18th-century England the Court of Chancery administered equity and the church courts handled the probate of wills. In the American colonies the governor and his council acted as a court of equity. For a time after independence, equity was suspect as a remnant of royal prerogative, but it has come to be generally applied by the same court as the regular law. Although U.S. common law is more flexible than English law, and the need for equity is less, important remedies have nevertheless been developed within the system. Probate, with a few exceptions, is usually a matter for the regular courts.

Federal and state judicial systems. State courts try 90 percent of all civil and criminal cases. Local magistrates may sit on county or district courts. One appeal is always given, and two levels of appeals exist in many states. The highest court is usually called the supreme court of the state, but this varies. In New York state, for example, the Supreme Court is a trial court, and the highest court is the Court of Appeals.

The Constitution of 1789 set up a federal Supreme Court, and the 1789 Judiciary Act provided for federal district courts and circuit courts. The plan for inferior courts has undergone changes from time to time, notably in 1891, when circuit courts of appeal were established, and in 1911, when the old circuit courts were abolished.

Most federal law is statutory and enforced by federal courts. Laws concerning tax, labour, securities regulations, admiralty, interstate commerce, antitrust, patent, and copyright matters fall into this category. By a decision of 1803, the Supreme Court became the ultimate authority for determining the conformity of all legislation with the federal Constitution, which guarantees many fundamental rights.

To ensure the fair treatment of out-of-state citizens or of corporations incorporated elsewhere, federal courts can try cases involving a diversity of citizenship. In such cases they act as if they were state courts, however, being bound by state statutes since 1842 and by state interpretations of common law and equity since 1938. Federal procedure is followed, but state rules on vital matters, such as statutes of limitations, are enforced.

Federal courts also try claims by and against the United States, such as cases undertaken to protect federal assets. In the absence of statutory provisions for such cases, a "federal common law" is applied.

Personal and property rights. The guarantees of due process of law given in Magna Carta in 1215 and the English Bill of Rights of 1689 are reflected in the first ten amendments to the federal Constitution, which were passed in 1791 and are known as the Bill of Rights. Since the passage of the Fourteenth Amendment in 1868, the rights of life, liberty, and property have been protected from deprivation by both the states and the federal government without due process of law; this has tended to shield private property from government regulation and private contracts from government interference. The use of property, however, is increasingly restricted by zoning laws and health and safety measures, and the acquisition of property for public purposes may be justified under the doctrine of eminent domain (power of the government to take private property for public use without the owner's consent upon payment of compensation).

The 1929 Depression was followed by the rejection by the Supreme Court of many welfare measures. Since 1937, however, the power of the Congress to regulate the economy under its authority to oversee interstate commerce has generally been upheld by the Supreme Court. State legislation is, as a rule, also held to be constitutional in this area. Minimum-wage laws and the right to collective bargaining in industry are recognized as well.

Human rights in the Supreme Court

Since the 1950s the emphasis in constitutionality cases has shifted to human rights. The requirement of "equal protection of the laws" and the Civil Rights Act of 1866 led to the ruling in 1954 that public schools must be racially integrated and to later rulings against using public funds for segregated private schools. The Federal Civil Rights Act of 1964 applies not only to official laws and actions but also to the conduct of private citizens. Thus, no discrimination on the basis of race, sex, religion, or national origin is allowed in places of public entertainment or resort or in employment practices by larger firms.

Since 1962 the Supreme Court has insisted on a regular redrawing of electoral districts to give each vote roughly the same value (seat reapportionment). It has also interpreted the constitutional prohibition of the establishment of a state religion to render school prayers and religious instruction illegal. In 1971 freedom of the press was held to justify *The New York Times* in publishing confidential political material.

COMPARISONS OF ENGLISH, AMERICAN, AND COMMONWEALTH LAW

The legal systems rooted in the English common law have diverged from their parent system so greatly over time that in many areas the legal approaches of common-law countries differ as much among themselves as they do with the civil-law countries. Indeed, England and the United States have so many legal differences that they are sometimes described as "two countries separated by a common law." The most striking differences are found in the area of public law: England has no written constitution and no judicial review, whereas every court in the United States possesses the power to pass judgment on the conformity of legislation and on other official actions to constitutional norms. Throughout the 20th century, many areas of U.S. law have been "constitutionalized" by the increasing exercise of judicial power. Other factors that account for much of the distinctiveness of public law in the United States are its complex federal system and its presidential, as distinct from parliamentary, form of government. In the area of private law, however, family resemblances among the common-law systems are much greater. Yet even there, despite broad basic similarities, the common-law countries have developed distinctive variations over time.

Personal law. The law of personal status (nationality, capacity, domicile, and so on) has been transformed by the advancement of the principle of equality of the sexes. In the area of divorce law, the intense legislative activity of the 1960s and 1970s left most common-law countries with systems of "mixed grounds" for divorce: one can obtain a divorce either for the fault of the other spouse or upon some no-fault ground such as separation or breakdown of the marriage. A minority of U.S. states have eliminated fault grounds entirely. The major differences among common-law systems appear in the legal treatment of the economic consequences of divorce: most common-law countries follow the English model that permits judges to use their own discretion in reallocating the property and income of the spouses in the way that seems fair; a minority of U.S. states adhere to the principle of equal rather than discretionary division of assets.

Property and succession. The basic principles of property and succession are much the same everywhere, but the newer countries have special laws on forests, mines, and water rights. In Australia, for example, the crown reserves all mineral rights to itself. The transfer of land in England is governed by a system of title registration. In Canada and the United States the separate deeds are recorded, and title insurance is widely used to protect the purchaser.

Succession on intestacy is broadly similar throughout common-law countries but varies everywhere in detail. The widow, for example, may get more in one country and the children more in another. All children of both sexes generally take equal shares. In regard to testate succession, nearly all U.S. states protect the surviving spouse against disinheritance by securing to him or her a fixed indefeasible share of the decedent's estate. In England, and most of the former Commonwealth countries, however, not only the spouse but also children and certain other dependents of the deceased are permitted to petition the court for discretionary financial provision out of an estate if, in the judgment of the court, the testator did not make reasonable provision for them.

In most U.S. states and some Canadian provinces there are homestead laws, which protect the family house or a certain minimum sum of money from the claims of creditors.

Extensions of tortious liability

Tort law. Tort law (*i.e.*, the law relating to private civil wrongs) is largely common law in England, Canada, and the United States. Several major reforms have been introduced along the same lines in different countries. Allowing claims by dependents of persons tortiously killed and removing the immunity of the crown or government or charitable institutions from tort claims provide examples. The liability of manufacturers to the ultimate consumer was first laid down by U.S. and then by English judges.

In the field of libel, U.S. practice is less strict than the English, and in the United States a public figure cannot sue for honest but unfair and untrue criticisms of his activities, whereas in England published facts must be true and comment fair. In some Australian states truth is not necessarily a defense to an action.

A notable U.S. tort is interference with privacy, examples being a stranger using one's photograph for advertising without permission, using "bugging" (*i.e.*, electronic eavesdropping devices) in one's home or searching it, or taking photographs of persons in embarrassing situations.

Contracts. Contract law is basically similar in the common-law countries. The most interesting difference relates to the question of enforcement of contracts by third parties who are not actually parties to the contract but who are persons for whose benefit the contract was made. English law excludes such rights, except in an occasional statute. The Indian Contract Code of 1872 generally allows it, as does U.S. state law.

English law still requires the use of a seal on a gratuitous contract (such as one agreeing to make a gift) but has largely repealed the laws requiring written evidence of ordinary contracts. Written evidence is often called for in the United States.

The various areas of special contracts, such as those applying to employment, sale of land, and agency, are broadly similar everywhere but are regulated by local legislation and by a wealth of labour legislation.

Criminal law and procedure. As regards criminal law and procedure, the substance of the law is much the same throughout the common-law countries. More important differences appear in the rules of criminal procedure. This rests in England on modern legislation, whereas the old procedure bore heavily on the accused. Accused persons may now testify at the trial or not, as they wish; they are entitled to legal counsel; and they are assisted out of public funds when they are accused of serious crimes and are unable to afford to pay the costs themselves.

Canada has a Dominion Criminal Code, which covers major crimes. It also has a Canadian Bill of Rights and provincial laws such as the Ontario Human Rights Code. India has an overriding Bill of Rights.

Developments in the United States are the most interesting. Criminal procedure has become a constitutional matter, with a kind of federal common law of criminal procedure overriding state law in many instances. Thus, "due process of law" under the Fourteenth Amendment to the federal Constitution and the Federal Rules of Criminal Procedure confer wide protection on accused persons—too wide, some think, for public safety.

Different approaches to the admissibility of evidence

English courts are reluctant to admit tape recordings unless supported by direct evidence of persons present, and this is generally the position taken in the United States, although, with the permission of a court, emergency wiretapping is permitted. English and U.S. law exclude confessions unless they are made freely and spontaneously. If evidence is found by unlawful means, such as by searching a house without a warrant, English law permits such evidence to be used, but U.S. law does not.

The main difference between English and U.S. safeguards is that English protections rest on statute or case law and may be changed by ordinary statute, whereas U.S. safeguards are constitutional and cannot be relaxed unless the Supreme Court later reverses its interpretation or the Constitution is amended.

The future of the common law. In the past the law performed the function of a referee in a free economy and was called in to apply generally accepted ideas of right and wrong to individual disputes. Today, law often forms an instrument of governmental policy or results from social pressures on the government. Law, therefore, is increasingly administrative.

Another tendency, and one that is likely to be reinforced, is an increasing reliance on statute law and codification as instruments of legal development. At one time the English Law Commission considered drafting a contract code, and the law of tort has been the subject of several statutes. When Britain entered the European Economic Community it was thought that there might be pressures to make English law more accessible by codifying it along the lines of the continental model. Harmonization of the laws of the member states, however, has not thus far required this. In the United States the legal sovereignty of the states impedes such a radical change, but uniform state laws are becoming more common.

In view of the general tendency in modern society of shielding the individual as fully as possible from the consequences of chance accidents, the judge-made law of tort may in time be replaced, as it has been in New Zealand, by a comprehensive system of official or private insurance, similar to the present compulsory third-party risk insurance available for motor vehicles. Public law is also gaining on private law in other fields—in real-property development, for example, public zoning or town-planning rules are already more important than the traditional restrictions imposed by individual neighbouring landowners. In family law, public-welfare laws on child care and adoption, pensions, and social security are often more important than the older private law based on the rights of spouses and children. (A.R.Ki./M.A.Gl.)

Comparison of civil law and common law

Between the 11th and 15th centuries the law of England was strongly influenced by Roman-law learning, and, in the 16th century, experts trained in Roman law were welcomed as administrators as much by the kings of England as by continental rulers. In contrast to the countries of the Continent, however, where justice was administered locally, the English judicial system had, as a result of the Norman Conquest, been centralized. There had grown up at the king's courts of Westminster a profession of practitioners expert in the law and procedure of the centralized court system, strongly organized and unwilling to yield its position, power, and income to a new group of specialists of Romanist learning. In its resistance to royal innovation, the organized bar allied itself with the parliamentary party in the great constitutional struggle of the 17th century. Thus, a reception of Roman law, such as had developed on the Continent, was prevented in England. At the same time a strong connection was established between the principles of constitutionalism and individual freedom, and the common law, which over time fostered the image of the common law as the legal system of freedom, in contrast to the civil law, in which the state is exalted over the individual. This image gained additional support from the fact that free political institutions were developed earlier and have been maintained more firmly in countries of the common law than in countries of the civil law.

Legal procedure in free societies

Political institutions are one thing, however, and techniques of dealing with civil litigation and criminal prosecution are another. Intimate connections exist, of course, between the two; a society is not free if civil cases are not handled impartially and if persons accused of crimes are not safeguarded against injustice. In both of these respects, however, neither of the two great legal systems lags behind the other. The ways of argumentation and procedure differ, but each of the two systems has developed its own guarantees and safeguards, and neither can be shown to be superior to the other. The view frequently found in England and the United States, that in civil-law criminal procedure the accused is presumed guilty until he has proved his innocence, is as unfounded as the view widely held on the Continent that trial by jury is tantamount to a lawless appeal to passion and emotion.

If one compares countries having firmly established institutions of constitutional government, such as the United Kingdom and the United States, to civil-law countries, such as Belgium, The Netherlands, Switzerland, the Federal Republic of Germany, or France, it appears that the protection of the individual against illegal actions by executive agencies is generally about the same. In some respects individual protection is even more elaborate in civil-law countries than in the United States and the United Kingdom. In general, it also may be said that it is less expensive for citizens on the Continent to seek legal protection of their private rights than it is in the common-law countries.

It is difficult to define what constitutes the real difference between common law and civil law. It would be erroneous simply to identify civil law with codified or even statutory law and common law with judge-made or case law. For one thing, the contrasts between the two systems existed long before the civil-law countries began to enact their codes. In addition, large parts of Anglo-American law are also contained in statutes or even codes, while in France, West Germany, and other civil-law countries parts of the law have never been reduced to statute at all but have been developed by the courts. Also, many of the statutes and code provisions have come to be overlaid by so many judicial opinions and interpretations that, in effect, they are dominated by judge-made law.

Role of judicial precedent

No essential difference can be found, either, in the role of judicial precedent. In theory, common-law courts are bound by precedent in the sense that once a legal question has been decided a certain way by a court, it must be decided in the same way by courts of inferior rank in the same jurisdiction until a higher court or the legislature sees fit to change the rule. Though higher courts are not "bound" by their previous decisions, they habitually follow them in the absence of a strong justification for overruling. In civil-law countries, on the other hand, courts are, in official theory, free to consider anew any legal question irrespective of how often it may have been determined before by other courts. In practice, however, common-law courts, especially those in the United States, have developed techniques for distinguishing new cases from older ones so that observable adherence to precedent is less a matter of strict obligation and more an acknowledgement of the importance of maintaining reasonable predictability in the law and of supporting the principle that like cases ought to be decided alike. Civil-law courts, for their part, have tended to follow precedent, not only for the sake of continuity and evenhandedness, but also in accord with the inclination of courts everywhere to save time and effort and to avoid reversal by a higher court.

The main difference between the systems consists of the ways in which the norms of the law are articulated and in which new rules are derived from older ones in novel cases. Though law cannot remain static, adaptation to new circumstances should be orderly and gradual so as not to unduly undermine predictability and stability in human affairs. In the common law, the role of adapting the law to changing conditions has traditionally belonged to judges. Because the administration of justice has been centralized, English judges have been well situated to assume this responsibility. In civil-law countries, where the multiplicity of courts has prevented this task from being performed by the judges, it has been professors who have taken the leading role.

Role of the professors

This difference in the identity of the influential lawmakers has had far-reaching consequences in the development of the law. Since courts must proceed from case to case, and cases arise in isolation and without prearrangement, a judge's opportunities for systematic theorizing are limited. Professors, by contrast, often deal with hypothetical cases as well as actual ones. They can develop comprehensive ideas and principles, and they are impelled for pedagogical reasons toward systematization and conceptualization. The civil law, as a professorial product, has thus tended to become more theoretical and consistent in its propositions and terminology than the judge-made common law, which has tended, for its part, to be closer to life and perhaps more detailed.

These traditional differences may diminish, however, for in the civil-law countries judicial power has increased with the national centralization of the administration of justice and with the assumption of a more active role by the judiciary, while in most of the common-law parts of the world the centralized English courts long ago lost their supremacy to a multiplicity of supreme courts in the United States and in the Commonwealth. The role of maintaining the internal coherence of the law is thus increasingly shared by the courts with that group of professionals—the professors—who traditionally have performed this function in the civil-law world. Gradual assimilation of the legal methods of the two great systems of law may thus well be expected.

Role of the adversary system

What is likely to continue, however, is the marked difference between the two great families of legal systems that exists in the field of procedure and, to some extent, in the personnel by whom justice is administered. All civil-law countries have adopted the adversary type of procedure that for centuries was peculiar to the common law, and they have abandoned the canonical procedure in which the evidence was presented to the judge in the form of a written record made up by a public officer, mostly in the absence of the parties. In modern adversary procedure, lawyers address their arguments directly to the court in both the civil-law and the common-law systems. Certain differences, however, are striking. In common-law systems the parties and their lawyers gather and present factual evidence in each case; in civil-law systems greater responsibility for investigation of facts is placed upon the judge, and it is generally the judge who plays the leading role in examining witnesses and who summons experts when needed. In general, the civil-law judge plays a much more active role in directing the course of a lawsuit than does his common-law counterpart.

Many differences between civil-law and common-law procedures and rules of evidence seem attributable to the absence of a jury in civil cases in the continental systems. During the 19th century, several civil-law countries experimented with the use of a jury for criminal cases. During the 20th century, however, the criminal jury has been abandoned in favour of a system having a mixed bench, on which professional, legally trained judges sit with a jury of laymen or lay judges to decide not only questions of fact (as in the common law) but also those of law.

(M.Rh./M.A.Gl.)

Variant and hybrid legal systems

Some Western legal systems cannot clearly be classified as belonging to the civil-law, common-law, or Socialist law traditions. Some developed in relative isolation from other systems, such as those in the Nordic countries; more often, variant legal systems arose from a confluence of influences, such as those that occurred when one colonial power succeeded another. Notable among these variants and hybrids are the Scandinavian, Roman-Dutch, and Scottish systems. Scandinavian law, though Germanic in origin and influenced by a revival of Roman law, is generally classified as sui generis and is set apart from the civil law by several distinctive features. Roman-Dutch law, an amalgam of Roman *jus commune* and the law of early modern Holland, survives in several former Dutch colonies in an altered form and in combination with several other legal influences. In Scotland, Roman law and common law have combined to form a hybrid system.

(M.A.Gl.)

SCANDINAVIAN LAW

Scandinavian law in medieval times constituted a separate and independent branch of early Germanic law and in modern times, in the form of codifications, became the basis of the legal systems of Norway, Denmark, Sweden, Iceland, and Finland.

Historical development of Scandinavian law. Before the Scandinavian states emerged as unified kingdoms in the 9th century, the several districts and provinces were virtually independent administratively and legally. Although social organization in the main was the same, and legal developments followed similar lines, there came into existence a number of separate legal systems, or "laws." Originally there were no written laws; the legal system consisted of customary law that was conserved, developed, and vindicated by the people themselves at the so-called *things,* or popular meetings of all free men. Between the 11th and 13th centuries the provincial customary laws were recorded in writing (invariably in the vernacular). These writings were most often private compilations but were occasionally instructions from the king. The best known laws of this period are the Gulathing's law (written in the 11th century, Norwegian); the law of Jutland (1241, Danish); and the laws of Uppland (1296) and Götaland (early 13th century), both Swedish. Other Scandinavian communities and states followed suit.

The early laws or codes did not have the character of civil codes as they are understood today. In addition to the subjects of private law (matrimony, inheritance, property, and contract), they contained constitutional and administrative law, criminal law, and laws of procedure. Ecclesiastical law was usually excluded and treated separately. In the main, the codes represented collections of customary law; influences from abroad were negligible except for some traces of canon law. Whereas the provincial laws, in common with other early Germanic laws, had tolerated and regulated blood feuds (setting up detailed tariffs for manslaughter and offenses against the body), the codes are, in several respects, more progressive. Thus, King Magnus' Swedish code (1350) abolished private vengeance, declaring that the king's officials should initiate criminal proceedings and provide for the punishment of wrongdoers. Furthermore, presumably under the influence of Christianity, legal provisions were introduced to assist paupers and the helpless. Rules concerning landed property (*e.g.,* the right of redemption belonging to the family) were markedly original.

Codification under Christian V

In 1380 Norway and Denmark were united under a common king (Olaf IV), but the two countries retained their separate laws. During the next 300 years, before the acquisition of absolute royal power by Frederick III (1660), supplementary laws were issued by the king in conjunction with an assembly of nobles. Finally, during the reign of Christian V, a comprehensive work of codification was accomplished, and the earlier and often obsolete law was replaced by Christian V's Danish Law (1683) and Norwegian Law (1687). The new codes were mainly based on the existing national laws of the two countries, and the influences of German, Roman, and canon laws were comparatively slight. Like the early codes, the newer codes consisted of public as well as private law and purported to treat exhaustively all more or less permanent legal rules and institutions. They were excellent codes for their times, drafted in a plain and popular style and inspired by respect for individual rights and the idea of equality before the law. The provisions of criminal law were relatively humane when compared with legislation in other European countries.

In Sweden a revised edition of the original code, issued by King Christopher (1442), was expressly confirmed by Charles IX (1608). The need for more modern legislation, however, made itself increasingly felt, and following the Danish-Norwegian example a royal commission was entrusted with the task of drafting a new code. The result, commonly called "the Law of 1734," was promulgated by Frederick I.

Finland, annexed by Sweden in the 13th century and made subject to Swedish law, came under the Swedish code of 1734, which was translated into Finnish as "Law of the Realm of Finland."

Modern Scandinavian law. The old codes have been all but completely displaced by modern parliamentary statutes. In Sweden the law of 1734 has been conserved as a formal framework. Elsewhere, plans for new and all-embracing codes are no longer entertained, but an extensive codification of important parts of the public and private law has taken place.

An interesting feature of Scandinavian law is the organized legislative cooperation that was begun in 1872 and has steadily increased in importance. In this way the Nordic states, including Iceland and Finland, have to a considerable degree obtained uniform legislation, especially regarding contracts and commerce, as well as in such fields of law as those concerned with family, the person, nationality, and extradition.

While conserving their national character, the Scandinavian legal systems have adopted certain conceptions of civil law (mainly German and French), chiefly through the influence of the law schools; commercial law and the laws of shipping and of companies, for example, conform more or less to common European patterns. Modern social welfare legislation, which has reached a high standard, also has strong international connections. Scandinavian law is pliable and close to life, less dogmatic than other European legal systems, and relatively free of formal rules and exigencies. Great attention is paid to rules and principles that have evolved in practice, especially in the courts. Much of the law is judge-made; and because the principle of stare decisis (*i.e.*, being bound by precedent) does not obtain, the courts have been free to meet the demands of changing social conditions. The extensive participation of laymen in both civil and criminal proceedings may have contributed in some measure to the pragmatic and flexible character of modern Scandinavian law. (F.Hi.)

ROMAN-DUTCH LAW

Roman-Dutch law is the system of law produced by the fusion of early modern Dutch law, chiefly of Germanic origin, and Roman, or civil, law. It existed in the Netherlands province of Holland from the 15th to the early 19th century and was carried by Dutch colonists to the Cape of Good Hope, where it became the foundation of modern South African law. It also influenced the legal systems of other countries that had once been Dutch colonies, such as Sri Lanka (formerly Ceylon) and Guyana.

Today Roman-Dutch law is in force throughout the Republic of South Africa and South West Africa/Namibia, and in Lesotho, Swaziland, Botswana, and Zimbabwe. In Sri Lanka it is present to a lesser degree, and in Guyana was from 1917 largely superseded by the common law of England. Reservation is made in favour of indigenous law and custom, so far as these are recognized; moreover the general law of these countries has in many respects departed from its original type.

Development of Roman-Dutch law in the Netherlands. In the 15th and 16th centuries the Roman law was "received" in the province of Holland (as it was sooner or later in the Netherlands generally), although general and local customs held their ground. These were based ultimately on Germanic tribal law—Frankish, Frisian, Saxon—supplemented by privileges and by-laws (*keuren*) and were themselves affected by an earlier infiltration of Roman law. The resulting mixed system, for which Simon van Leeuwen in 1652 invented the term "Roman-Dutch law," remained in force in the Netherlands until it was superseded in 1809 by the Code Napoléon, which in its turn in 1838 gave place to the Dutch civil code. The old law was also abrogated in the Dutch colonies. The Dutch civil code of 1838 has since been extensively revised.

There is, however, a third element in the Roman-Dutch system, namely the legislative acts of the Burgundian and Spanish periods, the most important of which were passed during the 16th century. Although a large quantity of legislation was later passed in the 17th and 18th centuries, it had little effect on the general character of the legal system. Roman-Dutch law can also be studied in collections of decided cases and of opinions (commonly termed *consultatien* or *advijsen*) and in the rich juristic literature of the system.

The commentaries of Grotius

The first attempt to reduce the Roman-Dutch civil law to a system was made by Hugo Grotius in his *Introduction to the Jurisprudence of Holland,* written while he was in prison in 1619–20 and published in 1631; this short treatise, a masterpiece of condensed exposition, remains a legal classic. Grotius' commentaries were followed by those of Johannes Voet and Simon van Groenewegen van der Made. Toward the end of the 18th century Dionysius Godefridus van der Keessel, professor at Leiden, lectured on the *jus hodiernum* ("law of today"), of which he published a summary in *Select Theses on the Laws of Holland and Zeeland . . .* (1800). The lectures, commonly known as the *Dictata,* still circulate as manuscript copies and have been cited in judgments by South African courts.

Survival and growth abroad of Roman-Dutch law. The law of the province of Holland was followed in the colonial empire, supplemented by local ordinances of the governors in council and, in the East Indies, by laws of the governors-general established at Batavia in Java (now Jakarta, Indon.). The ultimate legislative authority in the colonies was vested in the states general.

After the colonies passed to the British crown the old law underwent profound modifications, owing partly to changed social and economic conditions and partly to the incursion of rules and institutions derived from English common law.

The influence of English law (which was operative even during the period of the republics of the Transvaal and Orange Free State) has been most marked in criminal law and procedure, civil procedure, evidence, constitutional law, and, particularly, the commercial field of companies, bills of exchange, maritime law, and insurance. The law of tort or delict has also been considerably affected by English doctrines. On the other hand, the laws relating to property, persons, succession, and, to a lesser extent, contract still preserve their predominantly Roman-Dutch character. It is, for example, settled in both South Africa and Sri Lanka that "consideration" is not necessary for the validity of a contract.

The South Africa Act (1909) provided for the continuance of all laws in force in the several colonies at the establishment of the union until repealed by the Union Parliament or by the provincial councils within the sphere assigned to them. But thereafter, the Union Parliament and the appellate division of the Supreme Court of South Africa were active in consolidating, amending, and explaining the law and in making it more uniform. Many rules of the old law were pronounced obsolete by reason of disuse.

South African law

Modern South African law is a mixture of Roman-Dutch and English law. Constitutional law and administrative law have developed along English lines. The law of procedure and evidence is almost wholly English, as is most law relating to business associations and such areas as patents, trademarks, copyright, insurance, and maritime operations. On the other hand, criminal law is a combination of elements from Roman-Dutch and English common-law sources. In the law of succession, the rules governing the making of wills are English, whereas the substantive law of testamentary and intestate succession is largely Roman-Dutch. The law of persons and the law of property are almost purely Roman-Dutch, and the principles of the law of contract and of the law of delict are Roman-Dutch, only mildly influenced by common law.

(R.W.L./D.V.Cn.)

SCOTS LAW

At the union of the parliaments of England and Scotland in 1707, the legal systems of the two countries were very dissimilar. Scotland, mainly in the preceding century, had adopted as a guide much of the Roman law that had been developed by the jurists of Holland and France. But it is a fallacy to suppose that the law of Scotland is founded on the law of Rome: the Scots only turned to Roman, or civil, law when there was a gap in their own common or customary law. There is, however, a considerable infusion of civil law, not least in legal nomenclature and in the emphasis on principle rather than precedent. Perhaps the most important distinction is that Scotland, unlike England, did not separate the administration of equity and law.

Scots notion of equity

The Scottish conception of equity differs from the English system, which is parallel to the common law. The Scottish conception instead consists of a few fairly simple rules aimed at supplementing the law in order to prevent hardship. It also relegates certain remedies to the class of equitable remedies, of which the court has a large discretion to grant or withhold. The word equity in the law of Scotland has always retained its original meaning. The Scottish outlook upon this whole topic places Scots law clearly alongside the continental civil law and not the English system.

The merging of Scots and English law

Historical development of Scots law. The period following the union has been characterized by the merging of Scots and English law. One main cause of the merger is that much of the existing law of Scotland depends on statutes applicable to both countries. The House of Lords, consisting in its legal aspect until 1876 exclusively of English lawyers acting as the supreme court of appeal from Scotland, had a tendency to apply English law in Scottish appeals, and, in some cases, it ignored the distinction between its legislative and judicial functions. Another reason for the merging of systems is the influence of Scottish legal text writers, some of whom have tended to treat English law as though it were the law of their own country. The citation of English authorities in court has also had considerable effect.

Not surprisingly the most complete merger of the systems has occurred in the field of mercantile law. In other fields the systems are still widely separated.

Courts of law. The system of Scottish courts is completely different from that of the English and again is closer to the continental pattern. The supreme Scottish court (the House of Lords not being a native court) is the Court of Session, instituted by King James V in 1532, probably upon a French model. The court has two main functions. It has original jurisdiction in a very wide range of cases, which is exclusive in a few matters; in its appellate capacity it hears appeals (by reclaiming petition) from the nine Court of Session courts of first instance (called compendiously the Outer House), each presided over by a lord ordinary, and also from the sheriff courts. The appellate court (Inner House) sits in two divisions, the first and second, presided over, respectively, by the lord president of the Court of Session and the lord justice clerk. All the judges have the courtesy title of "lord" but are not on that account peers.

While the judges of the Court of Session are traditionally judges of both fact and law, in the early 19th century the civil jury was introduced, less because it was wanted in Scotland than because the House of Lords was weary of the great number of appeals it had to hear. Because the decision of a jury cannot in the ordinary sense be appealed, the House of Lords determined that caseloads would be drastically reduced by the change. From the Inner House appeal lies in many cases to the House of Lords by right and not, as in England, by leave. The right of audience in the Court of Session is possessed exclusively by members of the Faculty of Advocates (the Scottish Bar).

The sheriff court

The lower civil court is the sheriff court, which is an ancient court dating back to the 12th century. Scotland is divided into several sheriffdoms, each staffed by a sheriff-principal and a number of full-time sheriffs. Courts are held regularly in all the major towns of each sheriffdom. Sheriff courts have both civil and criminal jurisdiction. In civil cases, the sheriff normally makes decisions alone, although sometimes he is assisted by a jury of seven. In criminal cases, the accused is tried summarily or with a jury of 15. In civil jurisdiction appeal lies to the sheriff-principal and then to the Court of Session, or directly to the Court of Session; in criminal jurisdiction it lies to the High Court of Justiciary.

In addition to the sheriff court, there is the Summary Court, which hears minor pecuniary claims.

The Court of Session has absorbed the functions of certain ancient courts—the Court of Exchequer, the Admiralty Court, the Teind (or Tithe) Court, and the Commissary Court—which formerly dealt with questions of marriage law and executry, while the judges have by statute been given separate duties in a Lands Valuation Appeal Court, a Registration Appeal Court, and an Election Petition Court.

The Scottish Land Court, established in 1911, has jurisdiction in a wide range of matters relating to agriculture. Disputes between landlords and tenants of agricultural holdings may be brought before it by judicial process or, by agreement of the parties, in lieu of arbitration. It also deals with questions referred to it by the secretary of state for Scotland. (A.D.G./J.I.S./Ed.)

Socialist law systems

Socialist law derived from the system of public order devised by early Soviet leaders after the Russian Revolution in 1917. The Soviet model was subsequently emulated in those parts of the world where Communist parties became dominant, ranging from eastern Europe to Central and Southeast Asia and from the Caribbean to some parts of Africa. Socialist law took somewhat different forms in these countries because of differences in their prerevolutionary legal systems, which often were an amalgam of religious, customary, and received or imposed civil law. The immediate legal background of Socialist law in the Soviet Union and the eastern European Socialist countries was the Romano-Germanic system. For this reason and because Soviet law itself was much influenced by the civil-law tradition, some legal scholars have classified Socialist law as a subdivision of civil law. They have pointed out that, while the rules of substantive law in Socialist countries were influenced in varying degrees by the principles of Marxism and Leninism, the rules of civil and criminal procedure, the conceptual apparatus of the law, and the legal methodology were still essentially civilian. Guyana and Tanzania are notable exceptions because they are the only Socialist countries to have had a common-law background.

Other scholars, including Socialist jurists, have claimed an independent status for Socialist law. In their view, Socialist systems were distinguished from other legal systems by the influence of state ownership of the principal means of production, the special role that the Communist Party played in the legal system, the close relationship between the legal system and national economic planning, the denial of any distinction between public and private law, and, perhaps above all, a conception of the role of law as an instrument for restructuring society, shaping the new order, and dismantling the old. The claim that Socialist law constituted a separate legal family was recognized in practice after World War II when judges representing the Soviet legal system were elected to the International Court of Justice in accordance with requirements of the International Court's statute (article 9) that the court had to include judges who are representative of the main forms of civilization and of the principal legal systems.

THE HISTORICAL DEVELOPMENT OF SOVIET LAW

Guidelines provided by Marx and Engels

Since Karl Marx and Friedrich Engels never prescribed a particular system of law for a state organized to achieve the aims set forth in their *Communist Manifesto* of 1848, it was left for V.I. Lenin and his colleagues in the Russian Communist Party after the Bolshevik Revolution of 1917 to improvise a legal system. General guidelines provided by Marx and Engels stated that law should be regarded as an instrument of the state, not as a limitation upon those who make policy, and that it should enunciate rules of public order to facilitate the transition to Socialism and ultimately to Communism. It was to have two major tasks: (1) elimination of the political power of the bourgeoisie (the property-owning middle class) by depriving them of their ownership of productive resources; and (2) education of citizens in the disciplined pattern of life claimed to be requisite for the achievement of the social order they desired. Completion of these two tasks was expected to assure both the abundant production necessary to realize the aim of distribution according to need and also the self-discipline of citizens necessary to eliminate coercion as the method of preserving order. The police, the army, and courts would become unnecessary, and law would, in Marxist terminology, "wither away." Citizens would perform social obligations, expressed in morals and unsanctioned administrative regulations, because they believed them to be desirable, not because they feared punishment for violating them.

The new principles were embodied in the first constitution of the Russian Soviet Federated Socialist Republic (R.S.F.S.R.), promulgated on July 10, 1918 (article 9), as follows:

> The basic task placed during the present transitional moment on the constitution of the R.S.F.S.R. is the establishment of

the dictatorship of the city and village proletariat and of the poorest peasantry in the form of a powerful all-Russian Soviet authority with the objective of complete suppression of the *bourgeoisie*, the exploitation of man by man and the installation of socialism, under which there will be neither division into classes nor a State authority.

Early implementation. Lacking any precise pattern for a legal system designed to achieve the purposes stated in the first constitution, the new government issued only a few decrees designed to establish a framework for the new society and then set up a primitive institutional structure to enforce them. The decrees deprived private individuals of the ownership of land, banks, insurance companies, merchant fleets, and large-scale industry as an implementation of the policy of expropriation of the bourgeoisie. They also created restrictions on the employers of labour and secularized marriage and divorce. The enforcement instruments, called "people's courts," operated without benefit of professional prosecutors or a bar.

"Revolutionary conscience" as a guide

To permit regulation of the multitudinous social relationships for which no new law was prescribed, judges were directed at the beginning to apply Russian imperial laws, but only to the extent that they had not been revoked by the Revolution and were not contrary to the revolutionary conscience of the judges. Perhaps because Lenin and his colleagues were too busy to do otherwise, or perhaps because they preferred to let develop in practice a new legal order, no all-inclusive systematized body of law was prescribed until 1922. Local judges were guided during the first years, apart from the basic decrees indicated, only by suggestions issuing from the people's commissar of justice in the form of instructions, procedural manuals, and journal articles praising some court decisions as the proper application of revolutionary conscience and denouncing others as improper.

Political enemies of the Communists were not brought before the courts but were condemned by political bodies created at the same time as the people's courts and called revolutionary tribunals, or they were arrested and imprisoned without public hearings by the political police (Cheka). Lenin and his jurists professed that the treatment of opponents outside the regular court system was only a temporary measure. Nevertheless, the conduct of revolutionary tribunals and of the Cheka greatly influenced the course of Soviet law long after their formal abolition in 1922, for they stimulated a lack of respect for formality and strict adherence to law even among those charged with enforcing and administering the law.

After 1924 new agencies emerged within the People's Commissariat of Internal Affairs (NKVD), later the Committee for State Security (KGB), which enabled Lenin's successor, Joseph Stalin, to rid himself of opponents. Following Stalin's death in March 1953, the Communist Party's first secretary, Nikita S. Khrushchev, disclosed that many innocent persons had been convicted and sentenced to long prison terms and even to death by special boards of the secret police. The Soviet method of preserving order was revealed officially to have been dual, with courts on one side dealing publicly with nonpolitical offenses and social disputes, while, at the same time, administrative boards and the security police were secretly punishing Stalin's enemies.

Stalin's dual system

Codification. The New Economic Policy of 1921 reintroduced a strictly controlled private enterprise system in limited areas of the economy to speed reconstruction and to overcome the devastating effects of world and civil war. This reintroduction necessitated, in the Communist view, a stabilization of law to induce capitalists to invest in the country. Lenin's commissar of justice, Dmitry I. Kurskii, considered these measures retrogressive but necessary, while Nikolay V. Krylenko, later to become federal commissar of justice, called them the natural evolution of five years of Soviet law. Stabilization efforts resulted in the first systematic codification of Socialist law and led to the construction of a complex institutional framework of courts.

On Oct. 31, 1922, a judiciary act established within the Russian republic a three-tiered system of general courts with civil and criminal jurisdiction:

1. The name "people's court" was retained for the lowest level, but the institution was less primitive than that of 1917. Its bench comprised a full-time judge appointed annually and two lay judges selected from a panel of intelligent and politically trustworthy citizens, each serving for a few days. Lay judges shared responsibility with the professional judge in deciding issues of law and fact.
2. Appeals from this court were handled by a new provincial court that evolved from the Congress of People's Judges, which, under the prior rules, had gathered periodically in each province to survey the work of the local courts. The provincial courts also had jurisdiction to try offenses against the security of the regime and other serious civil and criminal cases. As an appellate court the provincial court consisted of a bench of three professional judges. As a trial court it resembled the people's court, except that the lay judges were selected from a more experienced panel of politically sound citizens.
3. The Supreme Court of the republic was created to coordinate policy in all provinces. Evolving from a control department established earlier in the Commissariat of Justice, this court was authorized not only to hear appeals from cases tried in provincial courts but also to discipline lower courts, to issue rulings interpreting the codes, and to try cases of an unusually important nature.

To provide specialized treatment for crimes relating to military matters and to the disruption of transport, then deemed critical to the success of the regime, special courts that had evolved earlier were continued by the judiciary act and subordinated to the Supreme Court. Following experiments that used nonprofessionals to conduct prosecution and defense, an office of public prosecutor and a college of defenders, who performed the functions of a bar, were established to aid the judges.

The substantive law and procedure applied by the new courts were established by codes enacted in 1922 and 1923; included were criminal, civil, family, land, and labour codes as well as codes of criminal and civil procedure. The drafters of these codes essentially followed patterns similar to those of the Romanist states of the European continent. The codes survived, with amendments, until the 1960s and '70s.

Patterns of legal institutions and the substantive and procedural law enacted by the R.S.F.S.R. were copied with little variation by the Soviet-type republics that emerged in peripheral regions of the old Russian Empire. The only change caused by the federation of the Soviet republics into the Union of Soviet Socialist Republics on Dec. 30, 1922, was the establishment, as a coordinator of all practice, of a Supreme Court of the U.S.S.R. The first federal constitution, adopted provisionally on July 6, 1923, and permanently on Jan. 31, 1924, extended to the federal government the authority to enact general principles of law that would be followed by the republics in maintaining their codes. While this development might have caused a change in the Soviet legal system, it did not; the first federal judiciary act of Oct. 29, 1924, confirmed the system of courts that already existed in each of the republics, and no change was made in substantive or procedural law. The sole innovation was the placing of the military and transport courts under the control of the Supreme Court; thus was created a self-contained system of federal courts, with lower branches functioning throughout the republics but beyond the reach of republic officials. A sharp line was thus drawn between courts that dealt with matters vital to the security of the regime and those concerned with social disputes.

Soviet federal system

The pattern established for Soviet courts and codes of law in 1922 and 1923 continued until Stalin's death, although the second federal constitution of Dec. 5, 1936, transferred authority to the federal level to enact codes of law to replace those of the republics. Although a second federal judiciary act was issued on Aug. 16, 1938, new codes were not enacted. On Feb. 11, 1957, the constitution was amended to restore the original relationship between individual republics and the federation, namely that they enact their own codes but conform to the general principles established federally. The first of the new general principles were enacted for criminal law and procedure on

Dec. 25, 1958, and for civil law and procedure on Dec. 8, 1961. Others later were enacted at intervals, and codes for the republics were subsequently revised. The major innovation of the 1958 enactments was the requirement that punishment be ordered only by a court in accordance with the rules of the procedural code. This provision was heralded by Soviet jurists as a means of preventing the return to extralegal procedures such as those exercised by the secret police during Stalin's dictatorship.

Applications to other Communist states. The application of the pattern of law that had evolved within the Soviet Union was first exported to other areas when Outer Mongolia declared the establishment of a "people's government" on July 11, 1921. Its provisions for courts and prosecutors were like those of the Soviet Union, and the subsequently adopted law on land use of Feb. 6, 1942, provided the same formula of state ownership of land allocated for use gratis and in perpetuity by the nomadic herdsmen for their privately owned cattle. A labour law of Feb. 14, 1941, followed the Soviet pattern, as did a social insurance law of June 22, 1942.

The extension of Soviet law

The major opportunity to install the Soviet legal system outside the Soviet Union came only after World War II, when Soviet-type governments were created in the European states that were occupied by the Soviet army. The system was also established in China, North Korea, and North Vietnam when military victories placed the Communists in power. In each country there was some variation, related in part to the degree of national economic development, from the Soviet pattern evolved in the Soviet Union.

Throughout all of the Communist states, law was given the functions previously established in the Soviet Union. Thus, the first constitution of the Hungarian People's Republic of Aug. 20, 1949, declared (article 41):

> Courts of the Hungarian People's Republic punish enemies of the working people, defend and secure the state, economic and social structure of the people's democracy, its offices and the rights of the toilers, educate the toilers in the spirit of observance of the rules of socialist intercourse.

Regulations governing the organization of the people's courts in the People's Republic of China, promulgated on Sept. 3, 1951, specified that courts were "to consolidate the people's democratic dictatorship, uphold the new democratic social order and safeguard the fruit of the people's revolution." The People's Republic of Bulgaria was even more outspoken than the Soviet Union in clarifying the function of law; in its first constitution, enacted on Dec. 4, 1947, it granted the right to citizens to create associations, but (article 87) only "if they are not directed against state and public order established by the present constitution," and then it stated:

> It is forbidden and will be punished by law to form organizations having as their purpose taking from the Bulgarian people the rights and freedoms won by the people's uprising of September 9, 1944, and guaranteed by the present constitution, or to limit these rights and freedoms, to place under threat the national independence and state sovereignty of the country or to propagate open or concealed fascist or antidemocratic ideology, or to facilitate imperialist aggression, and also to participate in these organizations.

The first constitution of the People's Republic of China, promulgated on Sept. 20, 1954, nearly five years after proclamation of the People's Republic on Oct. 1, 1949, fixed a system of courts that emulated the judicial system in the Soviet Union. There were to be lay assessors, full-time professional judges, Soviet-type prosecutors, and three levels of courts; judges were "independent and subject only to the law," as they had been declared to be in the U.S.S.R. constitution.

Judicial power in Chinese law

The Chinese government, among its first actions, had abrogated all legal codes of the previous government. A commission to draft new codes was established in 1950, but little major legislation was enacted in China until the 1970s. In the interim, the courts were guided by governmental decrees, a few basic statutes, and the program of the Communist Party.

The Socialist states in eastern Europe moved more rapidly toward codification than had China, and, in contrast, confirmed their prewar civil codes, although with modifications. Poland adopted a General Statute on Civil Law with 119 articles on July 18, 1950. Before the expulsion of the Yugoslav Communist Party from the Cominform on June 28, 1948, Yugoslavia, under the constitution of Jan. 31, 1946, had followed the Soviet legal pattern. Under Tito, Yugoslavia accepted the concept of a monopoly of power in the Communist Party and the necessity for state ownership of productive resources. Tito, however, refused to accept the dictates of the Soviet Union on every detail, especially on the centralization of industrial control and the speed of implementing collectivization. The Yugoslav government experimented with bringing workers into the managerial process of public corporations. The government also broadened the autonomy of local government councils. Yugoslav legal philosophers rejected the view that the state could not begin to wither away until full achievement of Communism; they wanted to start relaxing coercion immediately. Stalin's attempts to unseat Tito failed, and after Stalin's death his successors attempted a reconciliation with Tito by indicating an acceptance of the idea that there could be many roads to Socialism.

Changes after the death of Stalin

Post-Stalin developments. Stalin's death in 1953 and his subsequent vilification by the Communist Party of the Soviet Union at its 20th congress three years later opened a new era in the law of Communist states. The people's democracies sought to be relieved of Soviet tutelage. Uprisings in Hungary and Poland forced Soviet leaders to accept the evolution of divergent Socialist legal systems. Polish jurists, including Communists, demanded humanistic attitudes toward legal procedure, greater flexibility within the system, and an increased regard for the individual; other eastern European states followed their lead. In 1958 Soviet legal philosophers formally endorsed a new humanism and declared that it was always fundamental to Lenin's thought.

Only Chinese Communists resisted the trend toward strengthened legal procedures, denouncing Soviet policies of humanism and legality in an open letter published in 1963. After 1957 Chinese leaders abandoned their policy of copying Soviet legal patterns and ceased their efforts to draft codes of law, declaring that they wanted flexibility in court. They evolved a legal system seemingly inspired by relics of traditional Chinese attitudes as modified by their own guerrilla experience before 1949 and by their study of Soviet practices during the period before 1921.

During the Cultural Revolution, which lasted from 1966 to 1976, the Chinese political leadership adopted the attitude that law was dispensable and an obstacle to political progress. In the post-Mao period, a major reevaluation of the role of law in a Socialist state took place, and in 1978 a long-range plan was adopted to enact legislation, reestablish the legal profession, improve court procedures, expand legal education, and make legal consciousness a part of the popular culture. In accordance with this program, China adopted the Constitution of 1982 as well as several important statutes, including election laws, a criminal code, codes of criminal and civil procedure, a marriage law, tax laws, and several laws on the organization of government, the courts, and the procuracy.

Concurrently with revival of concern for humanistic values, most Communist states revised their constitutions and law codes to give formal support to their claim that they had passed beyond capitalism and arrived at Socialism. New civil codes designed for Socialist economies were introduced in Poland and Czechoslovakia in 1964, the latter incorporating terminology that departed sharply from Romanist tradition.

Several features, however, remained unchanged in the law of these states, even after the introduction of variations. Most notable was the continuation of Lenin's principle that after the Russian Revolution no basis remained for the Romanist division of law into public and private spheres. To Lenin the concern of the Soviet state with every detail of social intercourse required the state to maintain the right and opportunity of intervention in any matter at any time. Consequently, he held that in the Soviet system all law must be public; that is, it must reflect the state's vital concern with the social relationships governed by the law. This principle, sometimes called *dirigisme* by French scholars, remains a characteristic of Socialist law.

CHARACTER OF SOVIET LAW

Montesquieu's concept of the separation of powers as embodied in the Constitution of the United States was rejected by the founders of the Soviet legal system. The legislature was made supreme over the executive and the judiciary. In principle only the enactments of the legislature were a source of law.

In reality, however, the executive became a source of law because the Council of Ministers often acted without convening the legislature. The second federal constitution of the Soviet Union, adopted in 1936, attempted to stop this practice, which violated the constitution of 1924, but the executive again soon usurped legislative authority. Post-Stalin reformers again sought to restore the monopoly of lawmaking power to the legislature. In practice, however, it was not the full assembly of the legislature but its derivative body, the presidium, elected from its membership to legislate during intervals between full sessions, that created the day-to-day changes in law. Although subsequent ratification from the full body was sought, it was extremely difficult in practice for the legislature to revoke presidium actions that had already gone into effect. The presidium thus became the most important source of law, except for proclamations of state economic plans and budgets and for the enactment of general principles for republic codes.

Legislative domination

Communist Party orders, except for certain decrees issued jointly by the party and a state agency, were not in the technical sense sources binding on courts. Still, the party provided the initiative for legislative action. Some vital laws, such as those reforming agriculture or industry, were given preeminent status if they were signed by the party secretary as well as by the chairman of the presidium when they were published as law.

Having embraced the principle of legislative supremacy and rejected the notion that judges could make law, the Soviet Union accepted no concept of judicial precedent like that of the Anglo-American common law. Nevertheless, the Supreme Court of the U.S.S.R., during the exercise of its authority to enunciate principles of interpretation, proclaimed such far-reaching principles as to make them de facto sources of law. Even individual decisions of the Supreme Court, although lacking the binding force of precedent, were given careful attention by lower courts, since they were published as guides. The Soviet approach to precedent was, in practice, generally similar to that of Romano-Germanic systems.

Because the Supreme Court of the U.S.S.R. was subordinate to the legislature, it had no right to declare a law unconstitutional or to create any limitation upon the legislature or the executive on the ground that some higher or general principles of law had been violated. The legislature controlled the judiciary at all levels above the people's court by exercising appointive and recall power through the soviet operating at the level of the court. Only people's judges were elected by the people. Elections followed the pattern established by Stalin for all representative bodies in the Soviet Union: the ballot consisted of one candidate who was chosen by local professional groups with the guidance of the Communist Party members in the groups. In practice, judicial decisions had to generally agree with Communist Party policy or the judge risked recall.

The constitutional provision that "judges are independent and subject only to the law" (article 154) had to be understood as a means of denying local officials the right to intervene in the formation of an individual decision on personal grounds, but not as a means of prohibiting Communist Party intervention when such intervention was properly formulated in criticism of a line of decisions out of keeping with party wishes.

Judges were not entirely deprived of initiative by the rule of legislative supremacy or by the leading role played by the Communist Party. In both the civil and criminal codes adopted in 1922 there was authorization to depart from the rigid provisions of the code under special circumstances. The Civil Code opened with article 1, which read, "Civil rights are protected by law except in cases where they are exercised contrary to their social and economic purpose." The exception permitted judges to apply their concept of state policy, when it was not specifically stated, to achieve an appropriate result in an individual situation, but the decision created no rule of law for future cases. The article was used frequently to justify judicial refusal to protect the rights of capitalists during the period of reconstruction (1922–28), but after 1930 it fell into disuse. Soviet jurists argued that it should be deleted from the code as being unrelated to a stable society from which all capitalist elements already had been eradicated, but it was retained in the civil-law general principles of 1961 (article 5).

The Criminal Code of 1922, which underwent revision in 1926, permitted the judge to exercise discretion in withholding a conviction despite the fact that a crime had clearly been committed, provided that the case posed no social danger. The judge also had the power to punish a citizen who had committed an "apparent" crime even though it was not defined as such by the code; the formula for such judicial initiative was that the act be found socially dangerous in a manner analogous to another act specifically defined as a crime. This provision, which led to unpredictability in the application of the law, was eliminated from the criminal law by the general principles adopted in December 1958.

Judicial discretion under the Criminal Code

SUBSTANTIVE LEGAL PROVISIONS

Soviet jurists and their colleagues in eastern Europe and other Communist states claimed uniqueness for their system because of its constitutionally defined function of promoting Socialism. The end, in their view, transformed the means. While the novelty of the system was, therefore, primarily its philosophical base, the substantive law was revised to introduce some novel institutions.

Property. The *Communist Manifesto*'s interpretation of property ownership as the source of political power was accepted by all Communist parties as the guiding principle to be followed in their revisions of the law. Consequently, private ownership of productive wealth was abolished everywhere, to some degree, as soon as state officials felt competent to administer production efficiently. The federal constitutions of the Soviet Union (1936 and 1977) set forth the ideal, denying to private individuals the right to employ labour in productive enterprise and reaffirming Lenin's decree of Feb. 19, 1918, which nationalized all land.

The ideal achievement of the goal was not realized everywhere. Polish and Yugoslav Communists continued to permit peasants to own farmlands and also allowed private citizens to conduct service trades, although with limited numbers of employees. China has permitted some private ownership of factories to continue, in addition to ownership by collectives and by the state.

The various legal codes reflected the preponderant importance given to state-owned productive property. For example, severer penalties were prescribed by the criminal code for the theft of state property than of private property. In the Soviet Union labour codes were concerned solely with state employment. The provisions of civil codes related to sales, contracts, property damage, and inheritance governed only consumer goods in traditional Romanist legal fashion, while new forms had been developed to govern transactions that involved state property. Title to state property was subject to restrictions: no sales or leases were permitted, with limited exceptions for land allocated to peasant households that had been temporarily disrupted either by military service or by relocation for industrial employment.

Land had been declared state property only in the Soviet Union and the Mongolian People's Republic (constitution of 1924). In all Marxian Socialist states except Poland and Yugoslavia, land had increasingly been brought under the administration of cooperative-type associations structured on the model of the Soviet collective farm. Peasants had been induced, often under strong pressure, to assign to these associations the use of plots originally owned by or assigned to them. A variation introduced in China (1958) is called a "commune," and it unites both local government and land administration into one administrative unit

The ownership of land and personal property that supervises "brigades" (groups of villages) and their subordinate "teams" (individual villages). Since the late 1970s, however, the communes have been weakened or even abolished. There is a growing trend among peasants to contract for land independently and to sell their surplus products in a free market.

The Soviet model collective farm placed the ownership of machinery, tools, commonly used farm buildings, and crops in the cooperative; but member families were declared the owners of their family homestead, which also included hand tools and barnyard animals. Small plots of state-owned land were assigned to each family for private use as a household garden. Farm government was carried out by a membership assembly that elected managers who were usually introduced to the members by central government authorities. Remuneration was in the form of a share of the produce that was graduated according to the value of work performed, computed by an accounting unit called a "labour day."

Agricultural cooperatives, generally regarded as transitional structures designed to guide peasants toward the superior type of organization patterned on industry, enabled peasants to become farm employees and receive wages and social insurance benefits.

Nonproductive property and the tools of artisans were left to private ownership and were labeled "personal property" to distinguish them from income-producing "private property." Early efforts to equalize the ownership of consumer goods had been abandoned by Stalin in 1930 and egalitarianism denounced as a petty bourgeois utopian idea. The shortage of goods had forced this change in attitude, and wages had been adjusted to individual work performance as a means of stimulating production. This differentiation in wages had led progressively to the weakening of restraints on inheritance and also to the ownership of luxury items.

The federal constitutions of the Soviet Union entrenched the systems of property incentives by guaranteeing inheritance and the private ownership of savings accounts. To preclude reversion to capitalist practices, various restrictions were placed on the use of personal property. Thus, dwellings were limited in size to satisfy only family needs; sales of dwellings and apartments could occur only after long intervals; members of a family sharing a joint household could own only one dwelling and lease excess space only temporarily and at rentals not exceeding those charged for occupancy of state-owned premises. Property purchased with income gained illegally could be expropriated. Neglected property could be expropriated if repairs were not made following a warning.

Restrictions on personal property

The determination to eliminate income resulting from the investment of private funds had been weakened by various paramount state pressures. Thus, interest was paid on state savings bank accounts to stimulate deposits as a means of accumulating funds available for state investment and to reduce the threat of an inflation of currency. Investment in state bonds also was encouraged, but returns were not paid in conventional interest but rather as lottery winnings. Inheritance from abroad was paid to heirs within the Soviet Union.

Marxian Socialist states outside the Soviet Union established comparable restrictions, and Czechoslovakia went as far as requiring that the source of personal ownership be "honest"—*i.e.*, derived mainly from work for the benefit of society. China, under Mao Zedong, took the strict position, opposing extensive use of property incentives among individuals and espousing asceticism.

Communal ownership was reserved in most Marxian Socialist states for the distant future when abundance had been achieved, money was no longer used as a medium of exchange, and the state had "withered away," in accordance with the prophecy of Marx and Engels in the *Communist Manifesto*. Until that time, the state represented the community and therefore sought to enlarge its segment of productive property.

Yugoslav Communists rejected the Soviet model of state ownership of industry. After 1950 they used a system based on the "social ownership" of factories, under which the title to industrial property did not reside with the state but instead with the employees of each factory, who were organized into a "workers' council." Management was appointed not by a central ministry, as in other Marxian Socialist states, but by the workers' council, which acted in concert with the local government and an industry-wide chamber that represented all the workers' councils in the republic in the same branch of industry. The primary aim of the Yugoslav system was to overcome the evils thought to be inherent in a centralized bureaucracy.

Economic planning. Production was planned in all Marxian Socialist states, although to varying degrees and through different types of institutions. Stalin favoured detailed planning and the centralized decisions of the Council of Ministers, who were advised by the State Planning Commission (Gosplan). The Yugoslav opposition to a centralized bureaucracy placed an emphasis on planning by local government and workers' councils; these groups, however, were subject to influences exerted by the central government, which allocated major capital investment and established national goals.

Approaches to the planning of production

After 1957 Stalin's successors experimented with forms of decentralized decision making; regional economic councils (Sovnarkhoz) were established for industries not directly related to national defense in the hope of stimulating managerial initiative. Most centralized industrial ministries were abolished from 1957 to 1965, but Gosplan and specialized financial and technical institutions retained control over the allocation of key resources. With the abolition of the Sovnarkhoz and the reconstitution of industrial ministries, in 1965 the Soviet Union returned to centralization—but in a form differing from Stalin's model. Under the new organization the responsibilities of the authorities of the republics were heightened, and plant managers were directed to search out local sources of supplies and markets.

The encouragement of local initiative went further in some Marxian Socialist states: Czechoslovak specialists even proposed to dramatize the change in emphasis by declaring that the public corporations administering state-owned property were the "owners" of the buildings and tools rather than only the administrators. This extreme experimentation was terminated by the Soviet intervention in 1968, although the Yugoslav Communists continued to rely on their system of local ownership and direction subject to indicative plans having no force of law.

An interesting legal feature connected with production in all Marxian Socialist states was the system of contracts that were executed by plant managers for implementing production plans. These contracts prescribed quantity, design, technical specifications, price, and delivery date, and they used plan requirements as a basis for negotiation. After 1957 elements of consumer choice were also considered. These contracts were often used by permanent state tribunals, called State Arbitration, to settle disputes. The tribunals also could require a contract to be executed to implement the plan if the parties could not agree on terms. China has eschewed a permanent tribunal, opting for mediation through ad hoc committees established by local economic commissions that are linked to the planning apparatus.

Contracts as tools for economic planning

Labour. While production and distribution in the Soviet Union were subject to compulsory orders, labour remained free to move as it wished, although exceptions occurred during the emergency conditions of World War II. Although the Soviet labour supply was regulated by monetary and other inducements, it continued to be subject to the centralized determination of job classifications, hours, and wages.

The role of collective labour agreements changed throughout the years as the central authorities increasingly prescribed details of employment. The agreements fell into disuse in the mid-1930s but were restored after World War II to mobilize the labour force for production. Since job classifications, wages, and hours were established by law, the agreements concentrated on bettering local working conditions and on indicating the labour union's preferences regarding the use of bonuses that were collectively earned as incentives for exceeding planned production.

Since production was planned and, therefore, was by definition rational, strikes had no place in Marxian Socialist states. Remedies sought by management or labour unions against apparent injustices were supposed to take the form of petitions to higher authority for legislative or administrative rectification. Collective agreements were enforced primarily through moral suasion, but they were strengthened by provisions of the penal code that applied to state managers who violated the code, the collective agreement, or legal activity of labour unions, and to labour union officials who disrupted production so severely as to commit the crime of "wrecking." Work stoppages did occur in Socialist countries, however. Major strikes took place in Poland and Hungary in 1956 when Communist influence weakened, and reforms had to be instituted to enable the Communist Party to maintain power. In the 1980s, strikes in Poland and Yugoslavia became a significant de facto element of industrial relations.

The "right to work" had been established by all constitutions of Marxian Socialist states, but no system had implemented it by court orders that forced an employment office to provide a job. The concept had come to require only the elimination of discrimination on the basis of race, sex, age, or domicile in filling such jobs as existed.

Individual employment contracts reflected the statutory provisions that governed employment and its termination. Since there was no individual bargaining in establishing the terms of employment, written contracts were often omitted in practice. A grievance procedure under labour union supervision was the established means by which aggrieved parties appealed to the courts for review of managerial decisions concerning payment and dismissal. Labour that became redundant with the advance of automation was to be retrained for other employment by state centres.

The "workers' self-government" of Yugoslavia presented a variation in the theory of employment relationships since workers could be said to employ themselves, being both owners and workers. Their activity was regarded as a social duty rather than a labour relationship, but in practice standards of employment were established to which management was required to adhere. These standards were generally the same as in other Marxian Socialist states.

Social insurance. Social insurance was introduced in all Marxian Socialist states to protect employees who were injured on the job or elsewhere. No coverage was offered to collective farmers until 1966, or to housewives and children; the only means of recovery for these groups were suits based on the Civil Code's provisions on obligations. These rules conformed to traditional Romanist legal models with some variation. Thus, the first Russian Civil Code (1922) required no fault for recovery, but this provision became conventional through court practice and later legislation. Fault became necessary for liability, except in cases in which the instrument causing injury was classified as particularly hazardous. The list of such instruments included automobiles.

The social utility of negligence suits

Social insurance benefits originally were inadequate because of the impoverishment of the state treasury. Consequently, victims of accidents were encouraged to sue for the difference between social insurance payments and total losses. Soviet policymakers later concluded that it was desirable to continue this practice even when social insurance funds had been increased, the argument being that state managers would be prompted to exercise care because of the threat of suits, which would likely reduce profit margins and their eventual chances for promotion. Social insurance benefits thus continued to be set at levels lower than damages suffered.

Damages were generally limited to the loss of wages or anticipated wages for unemployed victims. Poland and Czechoslovakia also continued the ancient practice of allowing recovery for pain and suffering. No capitalization of damages was permitted, in part to avoid the transfer of large sums to irresponsible persons and in part to discourage an accumulation of wealth that could provoke capitalist speculation.

No personal liability insurance was made available to individuals in the Soviet Union, except to foreigners, in the belief that its availability would reduce the deterrent effect created by civil suits for accidents. Czechoslovak and Hungarian civil codes, however, retained pre-World War II provisions that created the opportunity of purchasing liability insurance. State insurance provided the state enterprises in all Marxian Socialist states with protection from suits by employees injured by hazardous equipment, unless charges of fault were made. The burden of proving absence of fault was on the defendant.

State liability

Liability for injury was attached to all state activities, even those not of an economic character, such as with ministries and local government, although until 1961 in the Soviet Union sovereign immunity for governmental functions had to be waived explicitly by statute to permit a suit. Thereafter, liability was extended to all types of state administration unless immunity was claimed by specific law. The Czechoslovak Civil Code (1964) established an approach to state liability by drawing no distinction between types of state activities in its rule of liability to suit.

A duty to protect state property (article 131) and to observe rules of Socialist conduct (article 130) was created by the second Soviet federal constitution (1936). It initiated judicial practice establishing an obligation of a state enterprise whose property had been protected to pay damages to a victim injured while attempting to perform his duty. This principle was written into the Fundamentals of Civil Law in 1961. No implementation was provided, however, for failure to rescue another in performance of the obligation to observe the rules of social conduct. Soviet authors concluded that this duty was moral and not legal, and violation could create no civil liability in favour of a citizen or his heirs if no rescue was attempted.

Artistic creation and invention. The legal rights that protect authors, artists, and inventors present problems in socialized economies unless such persons are employed by the state. Practice indicates that authors, artists, and inventors are most imaginative when they are unimpeded by state supervision. To provide the requisite freedom coupled with the stimulation of initiative, Soviet codifiers created a status for innovators that stood midway between Western concepts and those of the Socialist ideal, in which wages were paid only for production. This status was referred to as "a law of personal relationships rather than of property relationships."

Inventors in the Soviet Union were offered two alternatives: an "author's certificate" or a patent. The first was granted after proof of novelty was submitted to the state patent office. It qualified the inventor to receive payments computed on the savings that accrued to the state from the use of the invention, with the computation being based on the best of the first five years of use. Additional benefits included tax exemption on a portion of the receipts, priority in the allocation of scarce living space, and acceptance into a technical school. The patent form was traditional, except that restrictions on private enterprise limited its use to assignment to state enterprises. In practice only foreigners choose patents, and their patent rights were protected under an international convention that was ratified by the Soviet Union in 1965.

Although other Marxian Socialist states in eastern Europe also offered a form similar to the "author's certificate," patents remained in common use until 1963, when the trend toward the Soviet model became marked. All states, however, retained the patent form for local inventions that were likely to have appeal abroad so as not to discourage potential foreign licensees.

Protection of artists' rights

Authors and artists were provided a more conventional protection than inventors, but they also were deprived of some of the traditional rights of a copyright owner. They could not bargain for royalties or sell their work to foreign publishers except through a state trading enterprise. Royalties were paid in accordance with a state tariff that was based on the type and size of the edition. In addition to monetary rewards, which were payable for life and to heirs for a term of years, moral protection was offered in the form of the rights to control editorial changes and the use of the manuscript. Community interests were exerted against an individual author's will if the work was regarded as a national treasure. In such a case, it could be translated or published continuously after the first disclosure by the author. Royalties were paid, except when the language of translation was other than Russian—the reasoning behind this being that minority peoples needed culture, but they might not have been able to afford it. Nonmonetary inducements, including medals and honorary titles, were also used to stimulate innovation.

In 1973 the Soviet Union joined the Universal Copyright Convention, which then had a membership of more than 70 countries, including the United States. Several other eastern European Marxian Socialist states were also members of the convention, and some belonged to the Berne Convention.

Family law. Under the 1977 Soviet constitution, the family was placed under the protection of the state. This protection was manifested by the express constitutional obligation of the state to establish an extensive network of child-care institutions, to organize and improve services to families, and to provide allowances and benefits to families with children. Soviet family policy was closely allied with population and labour policy, aiming simultaneously at raising the birth rate while enabling mothers to remain in the labour force.

Equality of the sexes, a major aim of the Russian Revolution, was not understood as being inconsistent with the special protection offered in Socialist countries to women, especially to those with children. The 1977 Soviet constitution provided that women should have equal rights with men but that the state should ensure the exercise of these rights by protecting the health of women; by providing material support for mothers, which included paid maternity leaves and other benefits; and by aiding unmarried mothers.

The rights of women

Another goal of the Revolution in the area of family relationships was the secularization of marriage. Before 1926 Soviet family law presented little variation from western European models, but the 1926 code introduced unregistered marriage and divorce. Marriage could be recognized as legally constituted if cohabitation were proved orally. Divorce could be registered at the request of one party after the termination of marital relationships had been proved.

A trend toward the discouragement of promiscuity emerged in the mid-1930s after evidence linked increasing juvenile delinquency to the breakup of homes. A law enacted in 1936 required both parties to appear for divorce if the care and maintenance of children was an issue and sought to discourage repeated divorces by establishing a system of graduated fees for each succeeding divorce. The 1926 provisions, which accepted proof of marriage or divorce other than by registration, were revoked by law in 1944. Thereafter registration was compulsory to establish a marriage, and courts were given sole jurisdiction over divorce. Judges were required, however, to attempt reconciliation and to grant divorce only when they were convinced that the restoration of the family would be impossible.

Hostility to the system grew during the succeeding years until a compromise position was formulated and adopted by law in 1968. While the registration of a marriage remained a requirement, divorce could be registered without prior court decision if both parties consented and if there were no offspring. In other circumstances court proceedings were compulsory.

The compromise on marriage and family

While other Marxian Socialist states generally followed the Soviet model, all introduced variations to preserve national cultural patterns. None but China has adopted the early attitudes espoused in the Soviet Union concerning unrestrained divorce, and even China has discouraged the use of its simple procedure of granting divorce at the request of either party. All Marxian Socialist states sought to preserve the conjugal family against gratuitous decisions to separate. The formula established no precise grounds for divorce; instead it authorized divorce only in the event of a complete and permanent breakdown of the marital relationship. In practice, marriages have generally been considered broken if cohabitation has ceased. Children born out of wedlock have been accorded the same rights as those born legitimate.

The marital property regime in the Soviet Union was established after early experimentation with individually owned property. The system that was finally chosen in 1926 incorporated the concept of community of acquests; that is, property owned individually by the spouses before marriage remained their separate property, but property acquired during marriage was co-owned, unless it was received by one spouse as a gift from outside or by inheritance. The Soviet model was favoured by civil-law systems in general and was adopted elsewhere among the Marxian Socialist states.

Criminal law. Novel definitions of crime emerged in Marxian Socialist states to reflect Socialist principles. To the traditional crimes against the person and property, the Soviet Union added "economic crimes." This development occurred in 1932 and coincided with the termination of the private enterprise system in favour of a system of state monopoly of large-scale production and state planning. Under these altered priorities, merchandising became a crime.

The constitutional prohibition of all private employment of labour for purposes of production was reflected in the Criminal Code, which penalized employers for those practices. Even artisans employing no labour were forbidden to engage in various production activities unless their customers provided the raw materials. The leasing of unused space in small private dwellings at rates exceeding those established for tenants of state-owned apartment buildings was also a crime.

Fear of opposition to Communist Party leadership prompted the early introduction in the Soviet system of penalties for "counterrevolutionary" crimes. This phrase was replaced in later Socialist criminal codes by the term "crimes against the state." Although all legal systems punish attempts to subvert state authority, the definitions and practices under Stalin exceeded the norms of Western democracies in their vagueness of expression and in the severity of penalties they prescribed. After Stalin's death a decade ensued in which punishments for these crimes were lessened, but a new epoch was initiated in 1966 when authors were prosecuted for publishing allegorical manuscripts abroad that had been rejected by Soviet censors. The grounds given for these prosecutions were that the authors should have anticipated the hostile use to which their writings would be put by enemies of the Soviet Union.

Crimes against the state

To implement Marxist attitudes that describe religion as an intolerable superstition to be discouraged, the conduct of religious schools and of social activities related to religious worship were prohibited in the Soviet Union. Strong measures were adopted during the 1920s when the religious hierarchies opposed Communist programs; although these were relaxed during World War II to gain the support of peasants, basic policies remained unchanged. An order in 1966 prescribed punishment for religious activities that were deemed harmful to health, and in practice the law was interpreted broadly in order to hamper worship by minority sects.

While other Marxian Socialist states introduced less restrictive laws, their leaders sought to discourage religion. Chinese Communists expressed the belief that monks and priests should be reeducated. Eastern European leaders initially were militant in prosecuting clergy, but later they relaxed their vigilance as religious institutions became more tractable. Poland was relatively liberal with respect to religion and religious practices, and Yugoslavia even went so far as to sign a concordat with the Vatican in 1966.

Procedural law. Romanist influence was strong in the Soviet codes of criminal and civil procedure, which resembled the French codes in their main features, except that they allowed for more possibilities of control by the state. Soviet procedural law in turn influenced all of the other Socialist systems.

Emphasis in criminal procedure was placed on investigation before trial, as in French practice. The examiner had to establish more than the prima facie case required of the grand jury in the traditional common-law system. The suspect had to be given an opportunity to testify and to produce witnesses and evidence in his defense. The examiner was authorized to bring the case to trial only if he became convinced of the suspect's guilt, so that the public trial consisted primarily of verifying the examiner's work rather than hearing the defendant's case for the first time. Contrast with the French model was provided by the fact that the examiner was not a magistrate but a subordinate to the office of prosecutor, and until 1958 in the Soviet Union the defendant was not permitted to have an attorney during the preliminary investigation. Even thereafter, only juveniles and certain special classes of defendants, such as deaf and mute persons, were given an absolute right to an attorney.

The "inquisitorial" system

Trial procedure followed the French model closely in that the judges were instructed to establish their personal conviction of guilt. To do so they could go beyond the evidence that was presented by the prosecution and defense. The court could call witnesses, seek its own experts, examine material evidence, and visit places connected with the crime. In practice, the judges did most of the interrogating in the courtroom and relied on the prosecutor and the defense only to bring out points that could have been overlooked. No rule of evidence of any kind bound the court, except for the requirements that evidence be relevant, probative, and not unduly repetitious.

Civil procedure required the observance of rules of evidence only for certain formalities as established by the civil codes, such as for contracts or wills. Otherwise, the judges were as free in civil cases as they were in criminal matters to seek or hear what they felt they needed to establish personal conviction. As in certain Romano-Germanic legal systems, the prosecutor had the authority to intervene in civil cases when intervention was deemed necessary to protect broader social interests.

Appellate courts that found violations of procedural requirements or of substantive law or that found evidence unconvincing were required to remand for a new trial. Revision of a sentence or of a decision without a retrial in a court in the original jurisdiction was permitted only when the original court had no jurisdiction, where an amnesty had been improperly applied, or where there had been no basis for trial. A penalty could not be increased on the defendant's appeal.

Protest of decisions

A singular feature of Soviet procedure was the authority given to Supreme Court presidents and to the prosecutors of the republics and of the Soviet Union to protest a civil or criminal decision, after it had become final on appeal or in the absence of appeal. This procedure permitted the reopening of convictions or even of acquittals if the highest legal authorities found grave errors in their periodic audits of inferior-court activities. Prosecutors frequently requested such reopenings, even of convictions obtained by inferior prosecutors. This created an opportunity of petition to superior prosecutors or to court presidents after all other remedies had been exhausted. At the same time it gave state authorities a second chance, not given to the convict or to the party of a civil suit, to change an otherwise final decision.

Socialist law in perspective. The industrialization that had been characteristic of the 20th century required legal systems in all states to accept increasing state intervention in social relationships. Employers were no longer free to execute labour contracts as they wished; property owners were subject to increasing restraints on use (in some countries farms or dwellings could be lost if they were not used productively); nationalization or extensive regulation in banking, industry, and transport became widespread; compulsory insurance against injury to third parties became commonplace, as had social insurance against industrial accidents; and the secularization of marriage and divorce nearly everywhere prevented citizens from conducting their marital relationships solely within religious rules.

Socialist law, in reflecting the pervasive presence of the state in human affairs, appeared to be an extreme extension of the trends common to all modern legal systems. Yet, with its sense of historic mission to advance society toward Socialism and ultimately to Communism, and with its educational ambition to help create the "new Socialist man" purged of selfish bourgeois propensities, Socialist law was closer to the religious legal systems of the world than it was to the civil or common law. Thus, despite its many common features with the Romano-Germanic systems, the conclusion seems justified that Socialist law differed in more than degree from other legal systems and therefore required separate categorization.

(J.N.H./M.A.Gl./Ed.)

BIBLIOGRAPHY

Roman law: W.W. BUCKLAND, *A Text-book of Roman Law from Augustus to Justinian*, 3rd ed. rev. (1975), is the standard reference in English. Other general introductions are R.W. LEAGE, *Roman Private Law Founded on the Institutes of Gaius and Justinian*, 3rd ed. edited by A.M. PRICHARD (1962); M. KASER, *Roman Private Law*, 3rd ed., trans. by ROLF DANNENBRING (1980; originally published in German, 10th rev. ed., 1977); BARRY NICHOLAS, *An Introduction to Roman Law* (1962, reprinted 1975); and J.A.C. THOMAS, *Textbook of Roman Law* (1976). For a scholarly treatment of Roman law in the classical era, see F. SCHULZ, *Classical Roman Law* (1951, reprinted 1954). For civil procedure, see M. KASER, *Das römische Zivilprozessrecht* (1966); and LEOPOLD WENGER, *Institutes of the Roman Law of Civil Procedure*, rev. ed. (1955; originally published in German, 1925). See also TONY HONORÉ, *Emperors and Lawyers* (1981).

For the historical development of Roman law, see the scholarly and highly readable accounts by H.F. JOLOWICZ and BARRY NICHOLAS, *Historical Introduction to the Study of Roman Law*, 3rd ed. (1972); and HANS JULIUS WOLFF, *Roman Law: An Historical Introduction* (1951, reprinted 1981). LEOPOLD WENGER, *Die Quellen des römischen Rechts* (1953), is a detailed historical study that relates Roman legal development to that of the surrounding legal systems. For the early history, see C.W. WESTRUP, *Introduction to Early Roman Law*, 5 vol. (1934–54). For a bibliographical appendix, see WOLFGANG KUNKEL, *An Introduction to Roman Legal and Constitutional History*, 2nd ed. (1973, reprinted 1975; originally published in German, 6th ed., 1971).

For the reception of Roman law in Europe, see PAUL KOSCHAKER, *Europa und das römische Recht*, 4th ed. (1966); and the briefer account by PAUL VINOGRADOFF, *Roman Law in Medieval Europe*, 2nd ed. (1929, reissued 1968). An interesting study may be found in W.W. BUCKLAND and ARNOLD D. MCNAIR, *Roman Law and Common Law: A Comparison in Outline*, 2nd ed. rev. by F.H. LAWSON (1952, reprinted 1965). For classified references to modern literature, see A. BERGER, *Encyclopedic Dictionary of Roman Law* (1953, reprinted 1980). Periodical literature from 1800 is elaborately indexed by L. CAES and R. HENRION, *Collectio Bibliographica Operum ad Ius Romanum Pertinentium* (1949–78), issued in two series with supplements. The volumes of *Iura* (irregular) contain bibliographies of current literature, including articles and reviews.

Germanic law: For sources, see *A General Survey of Events, Sources, Persons, and Movements in Continental Legal History* (1912, reprinted 1968); for an exhaustive survey, see KARL VON AMIRA, *Germanisches Recht*, 4th ed. edited by KARL AUGUST ECKHARDT, vol. 1, *Rechtsdenkmäler* (1960); for discussion, see R. BUCHNER, *Die Rechtsquellen*, published as a supplement to W. WATTENBACH and W. LEVISON, *Deutschlands Geschichtsquellen im Mittelalter: Vorzeit und Karolinger*, vol. 2 (1953); and EDWARD JENKS, *Law and Politics in the Middle Ages: With a Synoptic Table of Sources*, 2nd ed. (1913, reprinted 1970). For substantive law, see HEINRICH BRUNNER, *Deutsche Rechtsgeschichte*, 2nd ed., vol. 1 (1906, reprinted 1961), the classic treatment; H. CONRAD, *Deutsche Rechtsgeschichte: Ein Lehrbuch*, vol. 1, *Frühzeit und Mittelalter*, 2nd ed. (1962, reprinted 1982); and C. VON SCHWERIN, *Grundzüge der deutschen Rechtsgeschichte*, 4th ed. prepared by HANS THIEME (1950). For very early law, see MARCO SCOVAZZI, *Le origini del diritto germanico: fonti, preistoria, diritto pubblico* (1957).

For Visigothic and Burgundian law, see E.A. THOMPSON, "The Barbarian Kingdoms in Gaul and Spain," *Nottingham Mediaeval Studies*, 7:3–33 (1963); and P.D. KING, *Law and Society in the Visigothic Kingdom* (1972). For Anglo-Saxon law, see F.W. MAITLAND, "The Laws of the Anglo-Saxons," *The Collected Papers of Frederic William Maitland*, ed. by H.A.L. FISHER, vol. 3, pp. 447–473 (1911, reprinted 1981); and H.G. RICHARDSON and G.O. SAYLES, *Law and Legislation from Aethelberht to Magna Carta* (1966). For north Germanic law, see L.B. ORFIELD, *The Growth of Scandinavian Law* (1953).

Modern comparative law: The world's legal systems are comparatively presented in KONRAD ZWEIGERT and HEIN KÖTZ, *An Introduction to Comparative Law*, 2 vol. (1977; originally published in German, 1969–71), which contains in its second volume comparative studies of selected topics of the law of contracts, torts, and restitution; and in RENÉ DAVID and JOHN E.C. BRIERLEY, *Major Legal Systems in the World Today: An Introduction to the Comparative Study of Law*, 3rd ed. rev. (1985; originally published in French, 1966). Both works contain extensive bibliographies. For a historical discussion

explaining the origins, the essential values, and the unifying features of the Western legal heritage, see HAROLD J. BERMAN, *Law and Revolution: The Formation of the Western Legal Tradition* (1983). JOHN P. DAWSON, *The Oracles of the Law* (1968, reprinted 1978), is a penetrating analysis of the methods of legal thought in the Anglo-American, French, and German systems, as developed through the different roles played in them by judges and scholars. The *International Encyclopedia of Comparative Law* (1971–), is a compendium of articles by an international group of scholars that compares legal systems on a wide variety of topics. Current materials and articles are published in the *American Journal of Comparative Law* (quarterly); and the *International and Comparative Law Quarterly*. A continuing list of all books and articles published in English since 1790 on the subject of civil law can be found in CHARLES SZLADITS (comp.), *A Bibliography on Foreign and Comparative Law* (1953–), with supplements in the *American Journal of Comparative Law*.

Civil law: ARTHUR TAYLOR VON MEHREN and JAMES RUSSELL GORDLEY, *The Civil Law System*, 2nd ed. (1977), is a rich collection of cases and other source materials from France and Germany. A useful introduction for general readers as well as for those with legal training is JOHN HENRY MERRYMAN, *The Civil Law Tradition: An Introduction to the Legal Systems of Western Europe and Latin America*, 2nd ed. (1985). Helpful guides to the legal systems of particular countries are: F.H. LAWSON, A.E. ANTON, and L. NEVILLE BROWN (eds.), *Introduction to French Law*, 3rd ed. (1967); OTTO KAHN-FREUND, CLAUDINE LÉVY, and BERNHARD RUDDEN, *A Source Book on French Law: System, Methods, Outlines of Contract*, 2nd ed. (1979); F.H. LAWSON, *A Common Lawyer Looks at the Civil Law* (1955, reprinted 1977), primarily concerned with French law; E.J. COHN, *Manual of German Law*, 2nd rev. ed., 2 vol. (1968–71), a book written for lawyers; NORBERT HORN, HEIN KÖTZ, and HANSI G. LESER, *German Private and Commercial Law: An Introduction*, trans. from German (1982); MAURO CAPPELLETTI, JOHN HENRY MERRYMAN, and JOSEPH M. PERILLO, *The Italian Legal System: An Introduction* (1967); and ARTHUR TAYLOR VON MEHREN (ed.), *Law in Japan: The Legal Order in a Changing Society* (1963).

Common law: A.K.R. KIRALFY, *The English Legal System*, 7th ed. (1984); and PHILIP S. JAMES, *Introduction to English Law*, 11th ed. (1985), are general outlines. There are numerous general historical works, such as EDWARD JENKS, *A Short History of English Law: From the Earliest Times to the End of the Year 1939*, 6th ed. (1949); A.K.R. KIRALFY, *Potter's Historical Introduction to English Law and Its Institutions*, 4th ed. (1958); THEODORE F.T. PLUCKNETT, *A Concise History of the Common Law*, 5th ed. (1956); J.H. BAKER, *An Introduction to English Legal History*, 2nd ed. (1979); and S.F.C. MILSOM, *Historical Foundations of the Common Law*, 2nd ed. (1981), a difficult but classic text.

For the United States, see E. ALLAN FARNSWORTH, *An Introduction to the Legal System of the United States*, 2nd ed. (1983), a broad study of the American legal system. A good historical treatment is LAWRENCE M. FRIEDMAN, *A History of American Law*, 2nd ed. (1985).

General studies of other common-law nations include G.W. PATON (ed.), *The Commonwealth of Australia: The Development of Its Laws and Constitution* (1952); BORA LASKIN, *The British Tradition in Canadian Law* (1969); E. MCWHINNEY (ed.), *Canadian Jurisprudence: The Civil Law and Common Law in Canada* (1958); and M.C. SETALVAD, *The Common Law in India*, 2nd ed. (1970).

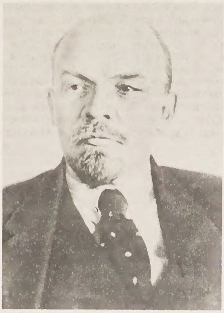
Socialist law: For a comprehensive survey of the Soviet legal system, see WILLIAM E. BUTLER, *Soviet Law* (1983). HAROLD J. BERMAN, *Justice in the U.S.S.R.: An Interpretation of Soviet Law*, rev. ed. (1963), is a study of Soviet law to the end of the Khrushchev era against the backdrop of Russian legal history and Marxist philosophy. The relationship of Soviet law to other Socialist legal systems is treated in JOHN N. HAZARD, *Communists and Their Law: A Search for the Common Core of the Legal Systems of the Marxian Socialist States* (1969). The effect of sociopolitical change on Soviet legal developments is discussed in JOHN N. HAZARD, *Managing Change in the U.S.S.R.* (1983). JOHN N. HAZARD, WILLIAM E. BUTLER, and PETER B. MAGGS, *The Soviet Legal System: The Law in the 1980's* (1984), presents original source materials and contains an extensive bibliography. The constitutions of Socialist legal systems are translated with introductions in WILLIAM B. SIMONS (ed.), *The Constitutions of the Communist World* (1980, reissued 1984).

(M.A.M./P.G.S./M.Rh./A.R.Ki./J.N.H./M.A.Gl.)

Lenin

Militant Marxist, founder of the Russian Communist Party (Bolsheviks), inspirer and leader of the Bolshevik Revolution, Lenin was the architect, builder, and first head of the Soviet state. He was the founder of the organization known as Comintern (Communist International) and posthumous source of "Leninism," the doctrine codified and conjoined with Marx's works by Lenin's successors to form Marxism-Leninism, which became the Communist worldview. If the Bolshevik Revolution is—as some people have called it—the most significant political event of the 20th century, then Lenin must for good or ill be regarded as the century's most significant political leader. Not only in the scholarly circles of the former Soviet Union but even among many non-Communist scholars, he has been regarded as both the greatest revolutionary leader and revolutionary statesman in history, as well as the greatest revolutionary thinker since Marx.

Tass—Sovfoto



Lenin, 1918.

EARLY LIFE

The making of a revolutionary. It is difficult to identify any particular events in his childhood that might prefigure his turn onto the path of a professional revolutionary. Vladimir Ilich Ulyanov was born on April 22 (April 10, old style), 1870, in Simbirsk, which from 1924 to 1991 was named Ulyanovsk in his honour. (He adopted the pseudonym Lenin in 1901 during his clandestine party work after exile in Siberia.) He was the third of six children born into a close-knit, happy family of highly educated and cultured parents. His mother was the daughter of a physician, while his father, though the son of a serf, became a schoolteacher and rose to the position of inspector of schools. Lenin, intellectually gifted, physically strong, and reared in a warm, loving home, early displayed a voracious passion for learning. He was graduated from high school ranking first in his class. He distinguished himself in Latin and Greek and seemed destined for the life of a classical scholar. When he was 16, nothing in Lenin indicated a future rebel, still less a professional revolutionary—except, perhaps, his turn to atheism. But, despite the comfortable circumstances of their upbringing, all five of the Ulyanov children who reached maturity joined the revolutionary movement. This was not an uncommon phenomenon in tsarist Russia, where even the highly educated and cultured intelligentsia were denied elementary civil and political rights.

As an adolescent Lenin suffered two blows that unquestionably influenced his subsequent decision to take the path of revolution. First, his father was threatened shortly before his untimely death with premature retirement by a reactionary government that had grown fearful of the spread of public education. Second, in 1887 his beloved eldest brother, Aleksandr, a student at the University of St. Petersburg (later renamed Leningrad State University), was hanged for conspiring with a revolutionary terrorist group that plotted to assassinate Emperor Alexander III. Suddenly, at age 17, Lenin became the male head of the family, which was now stigmatized as having reared a "state criminal."

Execution of Lenin's brother

Fortunately the income from his mother's pension and inheritance kept the family in comfortable circumstances, although it could not prevent the frequent imprisonment or exile of her children. Moreover, Lenin's high school principal (the father of Aleksandr Kerensky, who was later to lead the Provisional government deposed by Lenin's Bolsheviks in November [October, O.S.] 1917) did not turn his back on the "criminal's" family. He courageously wrote a character reference that smoothed Lenin's admission to a university.

In autumn 1887 Lenin enrolled in the faculty of law of the imperial Kazan University (later renamed Kazan [V.I. Lenin] State University), but within three months he was expelled from the school, having been accused of participating in an illegal student assembly. He was arrested and banished from Kazan to his grandfather's estate in the village of Kokushkino, where his older sister Anna had already been ordered by the police to reside. In the autumn of 1888, the authorities permitted him to return to Kazan but denied him readmission to the university. During this period of enforced idleness, he met exiled revolutionaries of the older generation and avidly read revolutionary political literature, especially Marx's *Das Kapital*. He became a Marxist in January 1889.

Formation of a revolutionary party. In May 1889 the Ulyanov family moved to Samara (known as Kuybyshev from 1935 to 1991). After much petitioning, Lenin was granted permission to take his law examinations. In November 1891 he passed his examinations, taking a first in all subjects, and was graduated with a first-class degree. After the police finally waived their political objections, Lenin was admitted to the bar and practiced law in Samara in 1892–93, his clients being mainly poor peasants and artisans. In his experience practicing law, he acquired an intense loathing for the class bias of the legal system and a lifelong revulsion for lawyers, even those who claimed to be Social-Democrats.

Law proved to be an extremely useful cover for a revolutionary activist. He moved to St. Petersburg (from 1914 to 1924 known as Petrograd; from 1924 to 1991 known as Leningrad) in August 1893 and, while working as a public defender, associated with revolutionary Marxist circles. In 1895 his comrades sent him abroad to make contact with Russian exiles in western Europe, especially with Russia's most commanding Marxist thinker, Georgy Plekhanov. Upon his return to Russia in 1895, Lenin and other Marxists, including L. Martov, the future leader of the Mensheviks, succeeded in unifying the Marxist groups of the capital in an organization known as the Union for the Struggle for the Liberation of the Working Class. The Union issued leaflets and proclamations on the workers' behalf, supported workers' strikes, and infiltrated workers' education classes to impart to them the rudiments of Marxism. In December 1895, the leaders of the Union were arrested. Lenin was jailed for 15 months and thereafter was sent into exile to Shushenskoye, in Siberia, for a term of three years. He was joined there in exile by his fiancée, Nadezhda Krupskaya, a Union member, whom he had met in the capital. They were married in Siberia, and

Exile in Siberia

she became Lenin's indispensable secretary and comrade. In exile they conducted clandestine party correspondence and collaborated (legally) on a Russian translation of Sidney and Beatrice Webb's *Industrial Democracy*.

Upon completing his term of Siberian exile in January 1900, Lenin left the country and was joined later by Krupskaya in Munich. His first major task abroad was to join Plekhanov, Martov, and three other editors in bringing out the newspaper *Iskra* ("The Spark"), which they hoped would unify the Russian Marxist groups that were scattered throughout Russia and western Europe into a cohesive Social-Democratic party.

Up to the point at which Lenin began working on *Iskra*, his writings had taken as their focus three problems: first, he had written a number of leaflets that aimed to shake the workers' traditional veneration of the tsar by showing them that their harsh life was caused, in part, by the support tsarism rendered the capitalists; second, he attacked those self-styled Marxists who urged Social-Democrats and workers to concentrate on wage and hour issues, leaving the political struggle for the present to the bourgeoisie; third, and ultimately most important, he addressed himself to the peasant question.

The principal obstacle to the acceptance of Marxism by many of the Russian intelligentsia was their adherence to the widespread belief of the Populists (Russian pre-Marxist radicals) that Marxism was inapplicable to peasant Russia, in which a proletariat (an industrial working class) was almost nonexistent. Russia, they believed, was immune to capitalism, owing to the circumstances of joint ownership of peasant land by the village commune. This view had been first attacked by Plekhanov in the 1880s. Plekhanov had argued that Russia had already entered the capitalist stage, looking for evidence to the rapid growth of industry. Despite the denials of the Populists, he claimed, the man of the future in Russia was indeed the proletarian, not the peasant. While attempting to apply the Marxist scheme of social development to Russia, Plekhanov had come to the conclusion that the revolution in Russia would have to pass through two discrete stages: first, a bourgeois revolution that would establish a democratic republic and full-blown capitalism; and second, a proletarian revolution after mature capitalism had generated a numerous proletariat that had attained a high level of political organization, socialist consciousness, and culture, enabling them to usher in full Socialism.

It was this set of principles that Lenin adhered to after he read Plekhanov's work in the late 1880s. But, almost immediately, Lenin went a step beyond his former mentor, especially with regard to the peasant question. In an attack on the Populists published in 1894, Lenin charged that, even if they realized their fondest dream and divided all the land among the peasant communes, the result would not be Socialism but rather capitalism spawned by a free market in agricultural produce. The "Socialism" put forth by the Populists would in practice favour the development of small-scale capitalism; hence the Populists were not Socialists but "petty bourgeois democrats." Lenin came to the conclusion that outside of Marxism, which aimed ultimately to abolish the market system as well as the private ownership of the means of production, there could be no Socialism.

The peasant question

Even while in exile in Siberia, Lenin had begun research on his investigation of the peasant question, which culminated in his magisterial *Development of Capitalism in Russia* (published legally in 1899). In this work, a study of Russian economics, he argued that capitalism was rapidly destroying the peasant commune. The peasantry constituted for the Populists a homogeneous social class, but Lenin claimed that the peasantry was in actuality rapidly stratifying into a well-off rural bourgeoisie, a middling peasantry, and an impoverished rural "proletariat and semi-proletariat." In this last group, which comprised half the peasant population, Lenin found an ally for the extremely small industrial proletariat in Russia.

Iskra's success in recruiting Russian intellectuals to Marxism led Lenin and his comrades to believe that the time was ripe to found a revolutionary Marxist party that would weld together all the disparate Marxist groups at home and abroad. An abortive First Congress, held in 1898 in Minsk, had failed to achieve this objective, for most of the delegates were arrested shortly after the congress. The organizing committee of the Second Congress decided to convene the congress in Brussels in 1903, but police pressure forced it to transfer to London.

The congressional sessions wore on for nearly three weeks, for no point appeared too trivial to debate. The main issues, nevertheless, quickly became plain: eligibility for membership and the character of party discipline; but, above all, the key questions centred around the relation between the party and the proletariat, for whom the party claimed to speak.

In his *What Is To Be Done?* (1902), Lenin totally rejected the standpoint that the proletariat was being driven spontaneously to revolutionary Socialism by capitalism and that the party's role should be to merely coordinate the struggle of the proletariat's diverse sections on a national and international scale. Capitalism, he contended, predisposed the workers to the acceptance of Socialism but did not spontaneously make them conscious Socialists. The proletariat by its own efforts in the everyday struggle against the capitalist could go so far as to achieve "trade-union consciousness." But the proletariat could not by its own efforts grasp that it would be possible to win complete emancipation only by overthrowing capitalism and building Socialism, unless the party from without infused it with Socialist consciousness.

The "vanguard of the proletariat"

In his *What Is To Be Done?* and in his other works dealing with party organization, Lenin articulated one of his most momentous political innovations, his theory of the party as the "vanguard of the proletariat." He conceived of the vanguard as a highly disciplined, centralized party that would work unremittingly to suffuse the proletariat with Socialist consciousness and serve as mentor, leader, and guide, constantly showing the proletariat where its true class interests lie.

At the Second Congress the *Iskra* group split, and Lenin found himself in a minority of opinion on this very issue. Nevertheless, he continued to develop his view of "the party of a new type," which was to be guided by "democratic centralism," or absolute party discipline. According to Lenin the party had to be a highly centralized body organized around a small, ideologically homogeneous, hardened core of experienced professional revolutionaries, who would be elected to the central committee by the party congress and who would lead a ramified hierarchy of lower party organizations that would enjoy the support and sympathy of the proletariat and all groups opposed to tsarism. "Give us an organization of revolutionaries," Lenin exclaimed, "and we will overturn Russia!"

Lenin spared no effort to build just this kind of party over the next 20 years, despite fierce attacks on his position by some of his closest comrades of the *Iskra* days, Plekhanov, Martov, and Leon Trotsky. They charged that his scheme of party organization and discipline tended toward "Jacobinism," suppression of free intraparty discussion, a dictatorship *over* the proletariat, not *of* the proletariat, and, finally, establishment of a one-man dictatorship.

The Bolsheviks and the Mensheviks

Lenin found himself in the minority in the early sessions of the Second Congress of what was then proclaimed to be the Russian Social-Democratic Workers' Party (RSDWP). But a walkout by a disgruntled group of Jewish Social-Democrats, the Bund, left Lenin with a slight majority. Consequently, the members of Lenin's adventitious majority were called Bolsheviks (majoritarians), and Martov's group were dubbed Mensheviks (minoritarians). The two groups fought each other ceaselessly within the same RSDWP and professed the same program until 1912, when Lenin made the split final at the Prague Conference of the Bolshevik Party.

CHALLENGES OF THE REVOLUTION OF 1905 AND WORLD WAR I

The differences between Lenin and the Mensheviks became sharper in the Revolution of 1905 and its aftermath, when Lenin moved to a distinctly original view on two issues: class alignments in the revolution and the character of the post-revolutionary regime.

The outbreak of the revolution, in January 1905, found Lenin abroad in Switzerland, and he did not return to Russia until November. Immediately Lenin set down a novel strategy. Both wings of the RSDWP, Bolshevik and Menshevik, adhered to Plekhanov's view of the revolution in two stages: first, a bourgeois revolution; second, a proletarian revolution (see above). But the Mensheviks argued that the bourgeois revolution must be led by the bourgeoisie, with whom the proletariat must ally itself in order to make the democratic revolution. This would bring the liberal bourgeoisie to full power, whereupon the RSDWP would act as the party of opposition. Lenin defiantly rejected this kind of alliance and post-revolutionary regime. Hitherto he had spoken of the need for the proletariat to win "hegemony" in the democratic revolution. Now he flatly declared that the proletariat was the driving force of the revolution and that its only reliable ally was the peasantry. The bourgeoisie he branded as hopelessly counterrevolutionary and too cowardly to make its own revolution. Thus, unlike the Mensheviks, Lenin henceforth banked on an alliance that would establish a "revolutionary democratic dictatorship of the proletariat and the peasantry."

Nor would the revolution necessarily stop at the first stage, the bourgeois revolution. If the Russian revolution should inspire the western European proletariat to make the Socialist revolution, for which industrial Europe was ripe, the Russian revolution might well pass over directly to the second stage, the Socialist revolution. Then, the Russian proletariat, supported by the rural proletariat and semi-proletariat at home and assisted by the triumphant industrial proletariat of the West, which had established its "dictatorship of the proletariat," could cut short the life-span of Russian capitalism.

Years of discouragement

After the defeat of the Revolution of 1905, the issue between Lenin and the Mensheviks was more clearly drawn than ever, despite efforts at reunion. But, forced again into exile from 1907 to 1917, Lenin found serious challenges to his policies not only from the Mensheviks but within his own faction as well. The combination of repression and modest reform effected by the tsarist regime led to a decline of party membership. Disillusionment and despair in the chances of successful revolution swept the dwindled party ranks, rent by controversies over tactics and philosophy. Attempts to unite the Bolshevik and Menshevik factions came to naught, all breaking on Lenin's intransigent insistence that his conditions for reunification be adopted. As one Menshevik opponent described Lenin: "There is no other man who is absorbed by the revolution twenty-four hours a day, who has no other thoughts but the thought of revolution, and who even when he sleeps, dreams of nothing but revolution." Placing revolution above party unity, Lenin would accept no unity compromise if he thought it might delay, not accelerate, revolution.

Desperately fighting to maintain the cohesion of the Bolsheviks against internal differences and the Mensheviks' growing strength at home, Lenin convened the Bolshevik Party Conference at Prague, in 1912, which split the RSDWP forever. Lenin proclaimed that the Bolsheviks were the RSDWP and that the Mensheviks were schismatics. Thereafter, each faction maintained its separate central committee, party apparatus, and press.

When war broke out, in August 1914, Socialist parties throughout Europe rallied behind their governments despite the resolutions of prewar congresses of the Second International obliging them to resist or even overthrow their respective governments if they plunged their countries into an imperialist war.

Denunciation of pro-war Socialists

After Lenin recovered from his initial disbelief in this "betrayal" of the International, he proclaimed a policy whose audacity stunned his own Bolshevik comrades. He denounced the pro-war Socialists as "social-chauvinists" who had betrayed the international working-class cause by support of a war that was imperialist on both sides. He pronounced the Second International as dead and appealed for the creation of a new, Third International composed of genuinely revolutionary Socialist parties. More immediately, revolutionary Socialists must work to "transform the imperialist war into civil war." The real enemy of the worker was not the worker in the opposite trench but the capitalist at home. Workers and soldiers should therefore turn their guns on their rulers and destroy the system that had plunged them into imperialist carnage.

Lenin's policy found few advocates in Russia or elsewhere in the first months of the war. Indeed, in the first flush of patriotic fervour, not a few Bolsheviks supported the war effort. Lenin and his closest comrades were left an isolated band swimming against the current.

Lenin succeeded in reaching neutral Switzerland in September 1914, there joining a small group of anti-war Bolshevik and Menshevik émigrés. The war virtually cut them off from all contact with Russia and with like-minded Socialists in other countries. Nevertheless, in 1915 and 1916, anti-war Socialists in various countries managed to hold two anti-war conferences in Zimmerwald and Kienthal, Switzerland. Lenin failed at both meetings to persuade his comrades to adopt his slogan: "transform the imperialist war into civil war!" They adopted instead the more moderate formula: "An immediate peace without annexations or indemnities and the right of the peoples to self-determination." Lenin consequently found his party a minority within the group of anti-war Socialists, who, in turn, constituted a small minority of the international Socialist movement compared with the pro-war Socialists.

Lenin's theory of imperialism

Undaunted, Lenin continued to hammer home his views on the war, confident that eventually he would win decisive support. In his *Imperialism, the Highest Stage of Capitalism* (1917), he set out to explain, first, the real causes of the war; second, why Socialists had abandoned internationalism for patriotism and supported the war; and third, why revolution alone could bring about a just, democratic peace.

War erupted, he wrote, because of the insatiable, expansionist character of imperialism, itself a product of monopoly finance capitalism. At the end of the 19th century, a handful of banks had come to dominate the advanced countries, which, by 1914, had in their respective empires brought the rest of the world under their direct or indirect controls. Amassing vast quantities of "surplus" capital, the giant banks found they could garner superprofits on investments in colonies and semi-colonies, and this intensified the race for empire among the great powers. By 1914, dissatisfied with the way the world had been shared out, rival coalitions of imperialists launched the war to bring about a redivision of the world at the expense of the other coalition. The war was therefore imperialist in its origins and aims and deserved the condemnation of genuine Socialists.

Socialist Party and trade-union leaders had rallied to support their respective imperialist governments because they represented the "labour aristocracy," the better paid workers who received a small share of the colonial "superprofits" the imperialists proffered them. "Bribed" by the imperialists, the "labour aristocracy" took the side of their paymasters in the imperialist war and betrayed the most exploited workers at home and the super-exploited in the colonies. The imperialists, Lenin contended, driven by an annexationist dynamic, could not conclude a just, lasting peace. Future wars were inevitable so long as imperialism existed; imperialism was inevitable so long as capitalism existed; only the overthrow of capitalism everywhere could end the imperialist war and prevent such wars in the future. First published in Russia in 1917, *Imperialism* to this day provides the instrument that Communists everywhere employ to evaluate major trends in the non-Communist world.

LEADERSHIP IN THE RUSSIAN REVOLUTION

By 1917 it seemed to Lenin that the war would never end and that the prospect of revolution was rapidly receding. But in the week of March 8–15, the starving, freezing, war-weary workers and soldiers of Petrograd (until 1914, St. Petersburg) succeeded in deposing the Tsar. Lenin and his closest lieutenants hastened home after the German authorities agreed to permit their passage through Germany to neutral Sweden. Berlin hoped that the return of anti-war Socialists to Russia would undermine the Russian war effort.

First return to Petrograd. Lenin arrived in Petrograd

The Provisional Government

on April 16, 1917, one month after the Tsar had been forced to abdicate. Out of the revolution was born the Provisional Government, formed by a group of leaders of the bourgeois liberal parties. This government's accession to power was made possible only by the assent of the Petrograd Soviet, a council of workers' deputies elected in the factories of the capital. Similar soviets of workers' deputies sprang up in all the major cities and towns throughout the country, as did soviets of soldiers' deputies and of peasants' deputies. Although the Petrograd Soviet had been the sole political power recognized by the revolutionary workers and soldiers in March 1917, its leaders had hastily turned full power over to the Provisional Government. The Petrograd Soviet was headed by a majority composed of Menshevik and Socialist Revolutionary (SR), or peasant party, leaders who regarded the March (February, O.S.) Revolution as bourgeois; hence, they believed that the new regime should be headed by leaders of the bourgeois parties.

On his return to Russia, Lenin electrified his own comrades, most of whom accepted the authority of the Provisional Government. Lenin called this government, despite its democratic pretensions, thoroughly imperialist and undeserving of support by Socialists. It was incapable of satisfying the most profound desires of the workers, soldiers, and peasants for immediate peace and division of landed estates among the peasants.

Only a soviet government—that is, direct rule by workers, soldiers, and peasants—could fulfill these demands. Therefore, he raised the battle cry, "All power to the Soviets!"—although the Bolsheviks still constituted a minority within the soviets and despite the manifest unwillingness of the Menshevik–SR majority to exercise such power. This introduced what Lenin called the period of "dual power." Under the leadership of "opportunist" Socialists, the soviets, the real power, had relinquished power to the Provisional Government, the nominal power in the land. The Bolsheviks, Lenin exhorted, must persuade the workers, peasants, and soldiers, temporarily deceived by the "opportunists," to retrieve state power for the soviets from the Provisional Government. This would constitute a second revolution. But, so long as the government did not suppress the revolutionary parties, this revolution could be achieved peacefully, since the Provisional Government existed only by the sufferance of the soviets.

Initially, Lenin's fellow Bolsheviks thought that he was temporarily disoriented by the complexity of the situation; moderate Socialists thought him mad. It required several weeks of sedulous persuasion by Lenin before he won the Bolshevik Party Central Committee to his view. The April Party Conference endorsed his program: the party must withhold support from the Provisional Government and win a majority in the soviets in favour of soviet power. A soviet government, once established, should begin immediate negotiations for a general peace on all fronts. The soviets should forthwith confiscate landlords' estates without compensation, nationalize all land, and divide it among the peasants. And the government should establish tight controls over privately owned industry to the benefit of labour.

Rise of the Bolsheviks

From March to September 1917, the Bolsheviks remained a minority in the soviets. By autumn, however, the Provisional Government (since July headed by the moderate Socialist Aleksandr Kerensky, who was supported by the moderate Socialist leadership of the soviets) had lost popular support. Increasing war-weariness and the breakdown of the economy overtaxed the patience of the workers, peasants, and soldiers, who demanded immediate and fundamental change. Lenin capitalized on the growing disillusionment of the people with Kerensky's ability and willingness to complete the revolution. Kerensky, in turn, claimed that only a freely elected constituent assembly would have the power to decide Russia's political future—but that must await the return of order. Meanwhile, Lenin and the party demanded peace, land, and bread—immediately, without further delay. The Bolshevik line won increasing support among the workers, soldiers, and peasants. By September they voted in a Bolshevik majority in the Petrograd Soviet and in the soviets of the major cities and towns throughout the country.

Decision to seize power. Lenin, who had gone underground in July after he had been accused as a "German agent" by Kerensky's government, now decided that the time was ripe to seize power. The party must immediately begin preparations for an armed uprising to depose the Provisional Government and transfer state power to the soviets, now headed by a Bolshevik majority.

Lenin's decision to establish soviet power derived from his belief that the proletarian revolution must smash the existing state machinery and introduce a "dictatorship of the proletariat"; that is, direct rule by the armed workers and peasants which would eventually "wither away" into a non-coercive, classless, stateless, Communist society. He expounded this view most trenchantly in his brochure *The State and Revolution,* written while he was still in hiding. The brochure, though never completed and often dismissed as Lenin's most "Utopian" work, nevertheless served as Lenin's doctrinal springboard to power.

Until 1917 all revolutionary Socialists rightly believed, Lenin wrote, that a parliamentary republic could serve a Socialist system as well as a capitalist. But the Russian Revolution had brought forth something new, the soviets. Created by workers, soldiers, and peasants and excluding the propertied classes, the soviets infinitely surpassed the most democratic of parliaments in democracy, because parliaments everywhere virtually excluded workers and peasants. The choice before Russia in early September 1917, as Lenin saw it, was either a soviet republic—a dictatorship of the propertyless majority—or a parliamentary republic—as he saw it, a dictatorship of the propertied minority.

Lenin therefore raised the slogan, "All power to the Soviets!", even though he had willingly conceded in the spring of 1917 that revolutionary Russia was the "freest of all the belligerent countries." To Lenin, however, the Provisional Government was merely a "dictatorship of the bourgeoisie" that kept Russia in the imperialist war. What is more, it had turned openly counterrevolutionary in the month of July when it accused the Bolshevik leaders of treason.

From late September, Lenin, a fugitive in Finland, sent a stream of articles and letters to Petrograd feverishly exhorting the Party Central Committee to organize an armed uprising without delay. The opportune moment might be lost. But for nearly a month Lenin's forceful urgings from afar were unsuccessful. As in April, Lenin again found himself in the party minority. He resorted to a desperate stratagem.

Around October 20, Lenin, in disguise and at considerable personal risk, slipped into Petrograd and attended a secret meeting of the Bolshevik Central Committee held on the evening of October 23. Not until after a heated 10-hour debate did he finally win a majority in favour of preparing an armed takeover. Now steps to enlist the support of soldiers and sailors and to train the Red Guards, the Bolshevik-led workers' militia, for an armed takeover proceeded openly under the guise of self-defense of the Petrograd Soviet. But preparations moved haltingly, because serious opposition to the fateful decision persisted in the Central Committee. Enthusiastically in accord with Lenin on the timeliness of an armed uprising, Trotsky led its preparation from his strategic position as newly elected chairman of the Petrograd Soviet. Lenin, now hiding in Petrograd and fearful of further procrastination, desperately pressed the Central Committee to fix an early date for the uprising. On the evening of November 6, he wrote a letter to the members of the Central Committee exhorting them to proceed that very evening to arrest the members of the Provisional Government. To delay would be "fatal." The Second All-Russian Congress of Soviets, scheduled to convene the next evening, should be placed before a *fait accompli.*

Overthrow of the Provisional Government

On November 7 and 8, the Bolshevik-led Red Guards and revolutionary soldiers and sailors, meeting only slight resistance, deposed the Provisional Government and proclaimed that state power had passed into the hands of the Soviets. By this time the Bolsheviks, with their allies among the Left SR's (dissidents who broke with the pro-Kerensky SR leaders), constituted an absolute majority of the Second All-Russian Congress of Soviets. The delegates therefore voted overwhelmingly to accept full power and elected Lenin as chairman of the Council of People's Commissars, the new Soviet Government, and approved his Peace Decree and Land Decree. Overnight, Lenin had vaulted from his hideout as a fugitive to head the Revolutionary government of the largest country in the world. Since his youth he had spent his life building a party that would win such a victory, and now at the age of 47 he and his party had triumphed. "It makes one's head spin," he confessed. But power neither intoxicated nor frightened Lenin; it cleared his head. Soberly, he steered the Soviet government toward the consolidation of its power and negotiations for peace.

Saving the Revolution. In both spheres, Lenin was plagued by breaks within the ranks of Bolshevik leaders. He reluctantly agreed with the right-wingers that it would be desirable to include the Menshevik and Right SR parties in a coalition government—but on Lenin's terms. They must above all accept the soviet form of government, not a parliamentary one; they refused. Only the Left SR's agreed, and several were included in the Soviet government. Likewise, when the freely elected Constituent Assembly met in January 1918, the Mensheviks and Right SR majority flatly rejected sovietism. Lenin without hesitation ordered the dispersal of the Constituent Assembly.

The Allies refused to recognize the Soviet government; consequently it entered alone into peace negotiations with the Central Powers (Germany and her allies Austro-Hungary and Turkey) at the town of Brest-Litovsk. They imposed ruinous conditions that would strip away from Soviet Russia the western tier of non-Russian nations of the old Russian Empire. Left Communists fanatically opposed acceptance and preached a revolutionary war, even if it imperilled the Soviet government. Lenin insisted that the terms, however ruinous and humiliating, must be accepted or he would resign from the government. He sensed that peace was the deepest yearning of the people; in any case, the shattered army could not raise effective resistance to the invader. Finally, in March 1918, after a still larger part had been carved out of old Russia by the enemy, Lenin succeeded in winning the Central Committee's acceptance of the Treaty of Brest-Litovsk. At last Russia was at peace.

Treaty of Brest-Litovsk

But Brest-Litovsk only intensified the determination of counterrevolutionary forces and the Allies who supported them to bring about the overthrow of the Soviet government. That determination hardened when, in 1918, Lenin's government repudiated repayments of all foreign loans obtained by the tsarist and Provisional governments and nationalized foreign properties in Russia without compensation. From 1918 to 1920 Russia was torn by a Civil War, which cost millions of lives and untold destruction. One of the earliest victims was Lenin himself. In August 1918 an assassin fired two bullets into Lenin as he left a factory in which he had just delivered a speech. Because of his robust constitution, he recovered rapidly.

The Soviet government faced tremendous odds. The anti-Soviet forces, or Whites, headed mainly by former tsarist generals and admirals, fought desperately to overthrow the Red regime. Moreover, the Whites were lavishly supplied by the Allies with materiel, money, and support troops that secured White bases. Yet, the Whites failed.

It was largely because of Lenin's inspired leadership that the Soviet government managed to survive against such military odds. He caused the formation and guided the strategy of the Workers' and Peasants' Red Army, commanded by Trotsky. Although the economy had collapsed, he managed to mobilize sufficient resources to sustain the Red Army and the industrial workers. But above all it was his political leadership that saved the day for the Soviets. By proclaiming the right of the peoples to self-determination, including the right to secession, he won the active sympathy, or at least the benevolent neutrality, of the non-Russian nationalities within Russia, because the Whites did not recognize that right. Indeed, his perceptive, skillful policy on the national question enabled Soviet Russia to avoid total disintegration and to remain a huge multinational state. By making the industrial workers the new privileged class, favoured in the distribution of rations, housing, and political power, he retained the loyalty of the proletariat. His championing of the peasants' demand that they take all the land from the gentry, church, and crown without compensation won over the peasants, without whose support the government could not survive.

Because of the breakdown of the economy, however, Lenin adopted a policy toward the peasant that threatened to destroy the Soviet government. Lacking funds or goods to exchange against grain needed to feed the Red Army and the towns, Lenin instituted a system of requisitioning grain surpluses without compensation. Many peasants resisted—at least until they experienced White "liberation." On the territories that the Whites won, they restored landed property to the previous owners and savagely punished the peasants who had dared seize the land. Despite the peasants' detestation of the Soviet's grain requisitioning, the peasants, when forced to choose between Reds and Whites, chose the Reds.

After the defeat of the Whites, the peasants no longer had to make that choice. They now totally refused to surrender their grain to the government. Threatened by mass peasant rebellion, Lenin called a retreat. In March 1921 the government introduced the New Economic Policy, which ended the system of grain requisitioning and permitted the peasant to sell his harvest on an open market. This constituted a partial retreat to capitalism.

The New Economic Policy

From the moment Lenin came to power, his abiding aims in international relations were twofold: to prevent the formation of an imperialist united front against Soviet Russia; but, even more important, to stimulate proletarian revolutions abroad.

In his first aim he largely succeeded. In 1924, shortly after his death, Soviet Russia had won de jure recognition of all the major world powers except the United States. But his greater hope of the formation of a world republic of soviets failed to materialize, and Soviet Russia was left isolated in hostile capitalist encirclement.

Formation of the Third International. To break this encirclement, he had called on revolutionaries to form Communist parties that would emulate the example of the Bolshevik Revolution in all countries. Dramatizing his break with the reformist Second International, in 1918 he had changed the name of the RSDWP to the Russian Communist Party (Bolsheviks), and in March 1919 he founded the Communist, or Third, International. This International accepted the affiliation only of parties that accepted its decisions as binding, imposed iron discipline, and made a clean break with the Second International. In sum, Lenin now held up the Russian Communist Party, the only party that had made a successful revolution, as the model for Communist parties in all countries. One result of this policy was to engender a split in the world labour movement between the adherents of the two internationals.

The Communist International scored its greatest success in the colonial world. By championing the rights of the peoples in the colonies and semi-colonies to self-determination and independence, the International won considerable sympathy for Communism. Lenin's policy in this question still reverberates through the world today. And it offers another example of Lenin's unique ability to find allies where revolutionaries had not found them before. By taking the side of the national liberation movements, Lenin could claim that the overwhelming majority of the world's population, then living under imperialist rule, as well as the European proletariat, were the natural allies of the Bolshevik Revolution.

Thus Lenin's revolutionary genius was not confined to his ability to divide his enemies; more important was his skill in finding allies and friends for the exiguous proletariat of Russia. First, he won the Russian peasants to the side of the proletariat. Second, while he did not win the workers to make successful Communist revolutions in the West, they did compel their governments to curtail armed intervention against the Bolshevik Revolution. Third, while the Asian revolutions barely stirred in his lifetime, they did strengthen the Soviet Communists in the belief that they were not alone in a hostile world.

By 1921 Lenin's government had crushed all opposition parties on the grounds that they had opposed or failed to support sufficiently the Soviet cause in the Civil War. Now that peace had come, Lenin believed that their opposition was more dangerous than ever, since the peasantry and even a large section of the working class had become disaffected with the Soviet regime. To repress opponents of Bolshevism, Lenin demanded the harshest measures, including "show" trials and frequent resort to the death penalty. Moreover, he insisted on even tighter control over dissent within the party. Lenin's insistence on merciless destruction of the opposition to the Bolshevik dictatorship subsequently led many observers to conclude that Lenin, though personally opposed to one-man rule, nevertheless unwittingly cleared the way for the rise of Joseph Stalin's dictatorship.

Suppression of dissent

By 1922 Lenin had become keenly aware that degeneration of the Soviet system and party was the greatest danger to the cause of Socialism in Russia. He found the party and Soviet state apparatus hopelessly entangled in red tape and incompetence. Even the agency headed by Stalin that was responsible for streamlining administration was, in fact, less efficient than the rest of the government. The Soviets of Workers' and Peasants' Deputies had been drained of all power, which had flowed to the centre. Most disturbing was the Great Russian chauvinism that leading Bolsheviks manifested toward the non-Russian nationalities in the reorganization of the state in which Stalin was playing a key role. Moreover, in April 1922 Stalin won appointment as general secretary of the party, in which post he was rapidly concentrating immense power in his hands. Soviet Russia in Lenin's last years could not have been more remote from the picture of Socialism he had portrayed in *State and Revolution.* Lenin strained every nerve to reverse these trends, which he regarded as antithetical to Socialism, and to replace Stalin.

Illness and death. In the spring of 1922, however, Lenin fell seriously ill. In April his doctors extracted from his neck one of the bullets he had received from the assassin's gun in August 1918. He recovered rapidly from the operation, but a month later he fell ill, partially paralyzed and unable to speak. In June he made a partial recovery and threw himself into the formation of the Union of Soviet Socialist Republics, the federal system of reorganization he favoured against Stalin's unitary scheme. However, in December he was again incapacitated by semiparalysis. Although no longer the active leader of the state and party, he did muster the strength to dictate several prescient articles and what is called his political "Testament," dictated to his secretary between Dec. 23, 1922, and Jan. 4, 1923, in which he expressed a great fear for the stability of the party under the leadership of disparate, forceful personalities such as Stalin and Trotsky. On March 10, 1923, another stroke deprived him of speech. His political activity came to an end. He suffered yet another stroke on the morning of Jan. 21, 1924, and died that evening in the village of Gorki (now known as Gorki Leninskiye), near Moscow.

The last year of Lenin's political life, when he fought to eradicate abuses of his Socialist ideals and the corruption of power, may well have been his greatest. Whether the history of the Soviet Union would have been fundamentally different had he survived beyond his 54th birthday, no one can say with certainty. (Al.Re.)

MAJOR WORKS

BOOKS: *Chto Takoye "Druzya Naroda," kak oni voyuyut protiv Sotsial-Demokratov?* (hectographic editions from 1894, printed from 1920; *What the "Friends of the People" Are, and How They Fight the Social-Democrats,* 1946); *Razvitiye kapitalizma v Rossi* (1899; augmented ed., 1908; *The Development of Capitalism in Russia,* 1956); *Chto delat?* (1902; *What Is To Be Done?,* 1929, 1933, 1950, 1963); *Shag vperyod, dva shaga nazad* (1904; *One Step Forward, Two Steps Back,* 1941); *Dve taktiki Sotsial-Demokraty v demokraticheskoy revolyutsi* (1905; *Two Tactics of Social-Democracy in the Democratic Revolution,* 1935, 1947); *Agrarny vopros i "Kritiki Marksa"* (1908, first chapters already published in periodicals from 1901); *Agrarnaya programma Sotsial-Demokratyi v pervoy russkoy revolyutsi 1905–1907 godov* (1908, 1917; *The Agrarian Programme of Social-Democracy in the First Russian Revolution,* 1954); *Imperializm, kak noveyshy etap kapitalizma* (1917, later retitled *Imperializm, kak vysshaya stadiya kapitalizma; Imperialism: The Latest Stage in the Development of Capitalism,* 1924; and *Imperialism, the Highest Stage of Capitalism,* 1933, 1939, 1947); *Gosudarstvo i revolyutsiya* (1917; *The State and Revolution,* 1919); *Proletarskaya revolyutsiya i renegat Kautsky* (1918; *The Proletarian Revolution and Kautsky the Renegade,* 1920); *Detskaya bolezn "levizny" v kommunizme* (1920; *"Left Wing" Communism: An Infantile Disorder,* 1920, 1934, 1961).

MARXIST PERIODICALS EDITED OR CONTROLLED BY LENIN: *Iskra,* 51 numbers (1900–03); *Zarya,* 4 numbers (1901–02); *Vperyod,* weekly, 18 numbers (1905); *Novaya Zhizn,* daily, 28 numbers (1905); *Proletary,* weekly, 26 numbers (1905); *Volna,* daily, 25 numbers (1906); *Vperyod,* daily, 17 numbers (1906); *Ekho,* daily, 14 numbers (1906); *Proletary,* 50 numbers (1906–09); *Sotsial-Demokrat,* 58 numbers (1908–17; wholly Leninist from 1911); *Zvezda,* at first weekly, then more frequent, 69 numbers (1910–12); *Rabochaya Gazeta,* 9 numbers (1910–12); *Prosveshcheniye,* monthly, 31 numbers (1911–14; and 1 final number, 1917); *Pravda,* daily (1912–14 and from 1917 with numerous changes of name 1913–14 and July–October 1917); *Kommunist,* 1 double number (1915); *Sbornik Sotsial-Demokrata,* 2 numbers (1916).

JOURNALISM AND PARTY THESES: "O zadachakh proletariata v dannoy revolyutsi" (*The April Theses,* 1951), "O dvoevlasti," and "Uroki revolyutsi" (*Lessons of the Revolution,* 1918), all in *Pravda,* 1917; "Ocherednyie zadachi Sovietskoy vlasti," *Pravda* (1918; *The Soviets at Work,* 1918; and *The Immediate Tasks of the Soviet Government,* 1951).

BIBLIOGRAPHY

The most complete collection of Lenin's works is *Polnoe sobranie sochinenii,* 5th ed., 55 vol. (1958–65), published in Moscow; it is supplemented by his *Biograficheskaia khronika, 1870–1924,* 13 vol. (1970–85), and *Leninskii sbornik,* 40 vol. in 31 (1924–85). The *Collected Works,* 45 vol. (1960–70), is a Soviet English translation of the 4th Russian edition of Lenin's works, enriched by editorial notes from the 5th edition. *Selected Works,* 3 vol. (1970–71), includes most of the works mentioned in this article and many more. Western publications of Lenin's works in English include *The Essentials of Lenin,* 2 vol. (1947, reprinted 1973), which follows the Soviet edition; and ROBERT C. TUCKER (ed.), *The Lenin Anthology* (1975), with interpretive comments.

Biographical and critical studies include ROBERT D. WARTH, *Lenin* (1973), an introductory study; ALFRED G. MEYER, *Leninism* (1957, reprinted 1986), an analysis of Lenin's political philosophy; DAVID SHUB, *Lenin,* rev. ed. (1966, reprinted 1977), a readable and informative biography by a contemporary; ADAM B. ULAM, *The Bolsheviks: The Intellectual and Political History of the Triumph of Communism in Russia* (1965, reissued 1973), a learned political biography; BERTRAM D. WOLFE, *Three Who Made a Revolution: A Biographical History,* 4th rev. ed. (1964, reissued 1984), a pioneering combined biography of Lenin, Trotsky, and Stalin, and *The Bridge and the Abyss: The Troubled Friendship of Maxim Gorky and V.I. Lenin* (1967, reprinted 1983); NADEZHDA KRUPSKAYA, *Reminiscences of Lenin* (1959; originally published in Russian, 1924), reticent, impersonal recollections by Lenin's widow; LEON TROTSKY, *Lenin: Notes for a Biographer* (1971; originally published in Russian, 1924), an appreciation of Lenin of the *Iskra* period and 1917–18, the periods of Trotsky's closest collaboration with Lenin, and *The Young Lenin,* trans. from Russian (1972); NIKOLAY VALENTINOV, *Encounters with Lenin* (1968; originally published in Russian, 1953), and *The Early Years of Lenin,* trans. from Russian (1969), revealing observations on Lenin's personality by a former associate; DIETRICH GEYER, *Lenin in der Russischen Sozialdemokratie* (1962), a scholarly study of Lenin and the origins of the Bolshevik-Menshevik split; LEONARD SCHAPIRO and PETER REDDAWAY (eds.), *Lenin: The Man, the Theorist, the Leader: A Reappraisal* (1967, reissued 1987), a collection of essays; ANGELICA BALABANOFF, *Impressions of Lenin* (1964), by the first secretary of the Communist International; and ROBERT SERVICE, *Lenin: A Political Life,* 3 vol. (1985–95).

Lenin is the subject of many historical studies, including RICHARD PIPES, *The Russian Revolution* (1990), and *Russia Under the Bolshevik Regime* (1993), the latter partly based on new archival sources; HAROLD SHUKMAN, *Lenin and the Russian Revolution* (1967, reissued 1977); HAROLD SHUKMAN and GEORGE KATKOV, *Lenin's Path to Power: Bolshevism and the Destiny of Russia* (1971); HELMUT GRUBER, *International Communism in the Era of Lenin: A Documentary History* (1967, reissued 1972), with interpretive essays; BRANKO LAZITCH and MILORAD M. DRACHKOVITCH, *Lenin and the Comintern* (1972); ALFRED ROSMER, *Lenin's Moscow* (1971, reissued 1987; also published as *Moscow Under Lenin,* 1972; originally published in French, 1953), an insider's account of the period 1920–24, exploring the role of the party in the international Communist movement; MICHAEL PEARSON, *The Sealed Train* (1975, reis-

sued 1989), an account of Lenin's associations with Germany and of the Russian Revolution; T.H. RIGBY, *Lenin's Government: Sovnarkom, 1917–1922* (1979); and HÉLÈNE CARRÈRE D'ENCAUSSE, *Lenin: Revolution and Power* (1982; originally published in French, 1979), a study of economic, social, political, and ideological issues. ESTHER KINGSTON-MANN, *Lenin and the Problem of Marxist Peasant Revolution* (1983), emphasizes the role of the peasants in the Russian Revolution. Lenin's influence on the Bolsheviks as well as their differences are addressed in ROBERT C. WILLIAMS, *The Other Bolsheviks: Lenin and His Critics, 1904–1914* (1986). R. CRAIG NATION, *War on War: Lenin, the Zimmerwald Left, and the Origins of Communist Internationalism* (1989), explores Lenin's part in the beginnings of Communism in the 20th century.

Interpretive studies of Lenin and Leninism include GEORG LUKÁCS, *Lenin: A Study on the Unity of His Thought* (1971; originally published in German, 1924), an evaluation by a Hungarian Marxist philosopher; DAVID LANE, *Leninism: A Sociological Interpretation* (1981); ALAIN BESANÇON, *The Rise of the Gulag: Intellectual Origins of Leninism* (1981; originally published in French, 1977); NEIL HARDING, *Lenin's Political Thought*, 2 vol. (1977–81); STANLEY W. PAGE, *The Geopolitics of Leninism* (1982), a critical look at Lenin's politics; and A.J. POLAN, *Lenin and the End of Politics* (1984), an analysis of Lenin's politics and influence. Official Soviet interpretation of Lenin's role is provided by BORIS N. PONOMAREV, *Lenin and the Revolutionary Process*, trans. from Russian (1980). NINA TUMARKIN, *Lenin Lives!: The Lenin Cult in Soviet Russia* (1983), explores the veneration of Lenin. Further references can be found in DAVID R. EGAN, MELINDA A. EGAN, and JULIE ANNE GENTHNER, *V.I. Lenin: An Annotated Bibliography of English-Language Sources to 1980* (1982).

Leonardo da Vinci

The unique fame that the Florentine artist and scientist Leonardo da Vinci enjoyed in his lifetime and that, filtered and purified by historical criticism, has remained undimmed to the present day is based on the equally unique universality of his spirit. Leonardo's universality is more than many-sidedness. True, at the time of the Renaissance and the period of Humanism, many-sidedness was a highly esteemed quality; but it was by no means rare. Many other good artists possessed it. Leonardo's universality, on the other hand, was a spiritual force, peculiarly his own, that generated in him an unlimited desire for knowledge and guided his thinking and behaviour. An artist by disposition and endowment, he found that his eyes were his main avenue to knowledge; to Leonardo, sight was man's highest sense organ because sight alone conveyed the facts of experience immediately, correctly, and with certainty. Hence, every phenomenon perceived became an object of knowledge. *Saper vedere* ("knowing how to see") became the great theme of his studies of man's works and nature's creations. His creativity reached out into every realm in which graphic representation is used: he was painter, sculptor, architect, and engineer. But he went even beyond that. His superb intellect, his unusual powers of observation, and his mastery of the art of drawing led him to the study of nature itself, which he pursued with method and penetrating logic—and in which his art and his science were equally revealed.

Alinari—Art Resource/EB Inc.

Leonardo, self-portrait, chalk drawing. In the Palazzo Reale, Turin, Italy.

LIFE AND WORKS

Early period: Florence. The illegitimate son of Ser Piero, a Florentine notary and landlord, Leonardo was born in 1452 on his father's family estate in Vinci, near Empoli. His mother, Caterina, was a young peasant woman who shortly thereafter married an artisan from that region. Not until his third and fourth marriages did Ser Piero's wives have children, the first one in 1476, when Leonardo was already an adult. Thus, Leonardo grew up in his father's house, where he was treated as a legitimate son and received the usual elementary education of that day: reading, writing, and arithmetic. As for Latin, the key language of traditional learning, Leonardo did not seriously study it until much later, when he acquired a working knowledge of it on his own. Not until he was 30 years old did he apply himself to higher mathematics—advanced geometry and arithmetic—which he studied with diligent tenacity; but here, too, he did not get much beyond the beginning stages.

Leonardo's artistic inclinations must have appeared early. When he was about 15, his father, who enjoyed a high reputation in the Florence community, apprenticed him to Andrea del Verrocchio. In Verrocchio's renowned workshop Leonardo received a many-sided training that included not only painting and sculpture but the technical-mechanical arts as well. He also worked in the next-door workshop of Antonio Pollaiuolo, where he was probably first drawn to the study of anatomy. In 1472 Leonardo was accepted in the painters' guild of Florence but remained five years more in his teacher's workshop. Then he worked independently in Florence until 1481. In the few extant works of this early period one may clearly trace the development of the artist's remarkable talent. Keenness of observation and creative imagination stand out. His early mastery is revealed in an angel and a segment of landscape executed by him in Verrocchio's painting the "Baptism of Christ" (Uffizi, Florence) and in two Annunciations (Uffizi, as well as the Louvre, Paris), both of them done in Verrocchio's workshop, as were the "Madonna with the Carnation," the "Madonna Benois," and the "Portrait of Ginevra de' Benci." This mastery reached its peak in two paintings that remained unfinished: "St. Jerome" and a large panel painting of "The Adoration of the Magi." In addition to these few paintings there are a great many superb pen and pencil drawings, in which Leonardo's mastery blazed new trails for this graphic art. Among the drawings are many technical sketches—for example, pumps, military weapons, mechanical apparatus—evidence of Leonardo's interest in and knowledge of technical matters at the outset of his career.

Evidence of early mastery

Unfolding of Leonardo's genius: first Milanese period (1482–99). In 1482 Leonardo entered the service of the Duke of Milan—a surprising step when one realizes that the 30-year-old artist had just received his first substantial commissions from his native city of Florence: the above-mentioned unfinished panel painting of "The Adoration of the Magi" for the monastery of S. Donato a Scopeto (1481) and an altar painting for the St. Bernard Chapel in the Palazzo della Signoria, which was never fulfilled. That he gave up both projects despite the commitments he had undertaken—not even starting on the second named—seems to indicate deeper reasons for his leaving Florence. It may have been that the rather sophisticated spirit of Neoplatonism prevailing in the Florence of the Medici went against the grain of his experience-oriented mind and that the more realistic academic atmosphere of Milan attracted him. Moreover, there was the fascination of Ludovico Sforza's brilliant court and the meaningful projects awaiting him there.

Leonardo spent 17 years in Milan, until Ludovico's fall from power in 1499. He was listed in the register of the royal household as *pictor et ingeniarius ducalis* ("painter and engineer of the duke"). Highly esteemed, Leonardo was constantly kept busy as a painter and sculptor and as a designer of court festivals. He was also frequently consulted as a technical adviser in the fields of architecture, fortifications, and military matters, and he served as a hydraulic and mechanical engineer.

In this phase of his life Leonardo's genius unfolded to the full, in all its versatility and creatively powerful artistic and scientific thought, achieving that quality of uniqueness that called forth the awe and astonished admiration of his contemporaries. At the same time, in the boundlessness of the goals he set himself, Leonardo's genius bore the mark of the unattainable so that, if one traces the outlines of his lifework as a whole, one is tempted to call it a grandiose "unfinished symphony."

The six paintings completed in the Milanese years

Painting and sculpture. As a painter Leonardo completed only six works in the 17 years in Milan: portraits of Cecilia Gallerani ("Lady with an Ermine") and a musician, an altar painting of "The Virgin of the Rocks" (two versions), a monumental wall painting of the "Last Supper" in the refectory of the monastery of Sta. Maria delle Grazie (1495–97), and the decorative ceiling painting of the Sala delle Asse in the Milan Castello Sforzesco (1498). Three other pictures that, according to old sources, Leonardo was commissioned to do have disappeared or were never done: a "Nativity" said to have belonged to Emperor Maximilian; a "Madonna" that Ludovico Sforza announced as a gift to the Hungarian king Matthias Corvinus; and the portrait of one of Ludovico's mistresses, Lucrezia Crivelli.

Also unfinished was a grandiose sculptural project that seems to have been the real reason Leonardo was invited to Milan: a monumental equestrian statue in bronze to be erected in honour of Francesco Sforza, the founder of the Sforza dynasty. Leonardo devoted 12 years—with interruptions—to this task. Many sketches of it exist, the most impressive ones discovered only in the mid-20th century, when two of Leonardo's notebooks came to light again in Madrid. They reveal the sublimity but also the almost unreal boldness of his conception. In 1493 the clay model of the horse was put on open display on the occasion of the marriage of Emperor Maximilian with Bianca Maria Sforza, and preparations were made to cast the colossal figure, which was to be 16 feet (five metres) high—double the size of Verrocchio's equestrian statue of Bartolomeo Colleoni! But, because of the imminent danger of war, the metal, ready to be poured, was used for cannon instead, and so the project came to a halt. Ludovico's fall in 1499 sealed the fate of this abortive undertaking, which was perhaps the grandest concept of a monument in the 15th century. The ravages of war left the clay model a heap of ruins.

As a master artist Leonardo maintained an extensive workshop in Milan, employing apprentices and students. The role of most of these associates is unclear. Their activity involves the question of Leonardo's so-called apocryphal works, in which the master collaborated with his assistants. Scholars have been unable to agree in their attributions of these works, which include such paintings as "La Belle Ferronnière" in the Louvre, the so-called "Lucrezia Crivelli" in the Pinacoteca Ambrosiana, Milan, and the "Madonna Litta" in the Hermitage, St. Petersburg (formerly Leningrad). Among Leonardo's pupils at this time were Giovanni Antonio Boltraffio, Ambrogio de Predis, Bernardino de' Conti, Francesco Napoletano, Andrea Solari, Marco d'Oggiono, and Salai.

Stimulus for Leonardo's scientific studies

Art and science: the notebooks. The Milan years also saw Leonardo's decided turn toward scientific studies. He began to pursue these systematically and with such intensity that they demanded more and more of his time and energy and developed into an independent realm of creative productivity. Within him there arose now a growing need to note and write down in literary form every one of his perceptions and experiences. It is a unique phenomenon in the history of art. Undoubtedly, the several treatises on art that appeared or were made available during those decades provided an external stimulus. Leon Battista Alberti's *De re aedificatoria* (*Ten Books on Architecture*) was first printed in 1485; Francesco di Giorgio's treatise on architecture was available in its first manuscript versions, and Leonardo had received a copy from the author as a gift. Moreover, Piero della Francesca in his *De prospectiva pingendi* ("On Perspective in Painting") had provided for his contemporaries a model text on the theory of perspective. Finally, there was the mathematician Lucas Pacioli, who had become an acquaintance of Leonardo's. In 1494 Pacioli published his *Summa de arithmetica geometria proportioni et proportionalità*, followed by his *Divina proportione* ("On Divine Proportion"), for which Leonardo drew figures of symmetrical bodies.

In this ambience Leonardo began to nourish the desire to write a theory of art of his own, and there arose in him the far-reaching concept of a "science of painting." Alberti and Piero della Francesca had already offered proof of the mathematical basis of painting in their analysis of the laws of perspective and proportion and thereby buttressed painting's claim to being a science. But Leonardo's claims went much further. Proceeding from the basic conviction that sight is the human being's most unerring sense organ, yielding immediate, accurate, and reliable data of experience, Leonardo—equating "seeing" with "perceiving"—arrived at a bold conclusion: the painter, doubly endowed with subtle powers of perception and the complete ability to pictorialize them, was the prime person qualified to achieve knowledge by observing and to reproduce that knowledge authentically in a pictorial manner. Hence, Leonardo conceived the staggering plan of observing all objects in the visible world, recognizing their form and structure, and pictorially describing them exactly as they are. Thus, drawing became the chief instrument of his didactic method.

In the years between 1490 and 1495 the great program of Leonardo the writer (author of treatises) began. In it, four main themes, which were to occupy him for the rest of his life, could be discerned and gradually took shape: a treatise on painting, a treatise on architecture, a book on the elements of mechanics, and a broadly outlined work on human anatomy. His geophysical, botanical, hydrological, and aerological researches also belong to this period and constitute parts of the "visible cosmology" that loomed before Leonardo as a distant goal. Against speculative book knowledge, which he scorned, he set irrefutable facts gained from experience—from *saper vedere*.

The notebooks

All these studies and sketches were written down in Leonardo's notebooks and on individual sheets of paper. Altogether they add up to thousands of closely written pages abundantly illustrated with sketches—the most voluminous literary legacy any painter has ever left behind. Of more than 40 codices mentioned in the older sources—often, of course, rather inaccurately—21 have survived; these in turn sometimes contain notebooks originally separate and now bound together so that 31 in all have been preserved. To these should be added several large bundles of documents: an omnibus volume in the Biblioteca Ambrosiana, Milan, called Codex Atlanticus because of its size, was collected by the sculptor Pompeo Leoni at the end of the 16th century; its sister volume, after a roundabout journey, fell into the possession of the English crown and was placed in the Royal Library, Windsor Castle. Finally there is the Arundel Manuscript (British Museum, MS. 263), which contains a number of Leonardo's fascicles on various themes.

It was during his years in Milan that Leonardo began the earliest of these notebooks. He would first make quick sketches of his observations on loose sheets or on tiny paper pads he kept in his belt; then he would arrange them according to theme and enter them in order in the notebook. Surviving are a first collection of material for the painting treatise (MSS. A and B in the Institut de France, Paris), a model book of sketches for sacred and profane architecture (MS. B, Institut de France, Paris), the treatise on elementary theory of mechanics (MS. 8937, Biblioteca Nacional, Madrid), and the first sections of a treatise on the human body (Anatomical MS. B; Windsor Castle, Royal Library).

Two special features make Leonardo's notes and sketches unusual: his use of mirror writing and the relationship between word and picture.

Leonardo was left-handed; so mirror writing came easily and naturally to him. It should not be looked upon as a secret handwriting. Though somewhat unusual, his script can be read clearly and without difficulty with the help of a mirror—as his contemporaries testified. But the fact that Leonardo used mirror writing throughout, even in his fair copies drawn up with painstaking calligraphy, forces one to conclude that, although he constantly addressed an imaginary reader in his writings, he never felt the need to achieve easy communication by using conventional handwriting. Yet occasional examples of normal handwriting (drafts of letters, notes, and comments to be submitted to third parties) show that Leonardo was completely at home in it. In the overwhelming majority of his notes in mirror writing, therefore, one gets the strong impression of "monologues in writing." Finally, then, his writings must be interpreted as preliminary stages of works destined for eventual publication, which Leonardo never got around to completing. In a sentence in the margin of one of his late anatomy sketches, he implores his followers to see that his works are printed.

Function of Leonardo's illustration vis-à-vis text

The second unusual feature in Leonardo's writings is the new function given to illustration vis-à-vis the text. Leonardo strove passionately for a language that was clear yet expressive. The vividness and wealth of his vocabulary were the result of intense self-study and represented a significant contribution to the evolution of scientific prose in the Italian vernacular. On the other hand, in his teaching method Leonardo gave absolute precedence to the illustration over the written word; hence, the drawing does not illustrate the text; rather, the text serves to explain the picture. In formulating his own principle of graphic representation—which he himself called *dimostrazione* ("demonstrations")—Leonardo was a precursor of modern scientific illustration.

Thus, during Leonardo's years in Milan the two "action fields"—the artistic and the scientific—developed and shaped his future creativity. It was a kind of "creative dualism," with mutual encouragement but also mutual pressure from each field.

Second Florentine period (1500–06). In December 1499 or at the latest January 1500—three months after the victorious entry of the French into Milan—Leonardo left that city in the company of Lucas Pacioli. He stopped first at Mantua, where, in February 1500, he drew a portrait of his hostess, Marchioness Isabella d'Este, and then proceeded to Venice (in March), where the Signoria (governing council) sought his advice on how to ward off a threatened Turkish incursion in Friuli. Leonardo recommended that they prepare to flood the menaced region. From Venice he returned to Florence, where, after a long absence, he was received with acclaim and honoured as a renowned native son. In that same year he was appointed an architectural expert to a committee investigating damages to the foundation and structure of the church of S. Francesco al Monte. A guest of the Servite order in the cloister of SS. Annunziata, Leonardo began there a cartoon for a painting of the "Virgin and Child with St. Anne," the composition of which won admiration from artists and art lovers of the city. He also painted (1501) a "Madonna with the Yarn-Winder," which has survived only in copies and which he probably never finished. Mathematical studies seem to have kept him away from his painting activity much of the time, or so Isabella d'Este, who sought in vain to obtain a painting done by him, was informed by Fra Pietro Nuvolaria, her representative in Florence.

Service with Cesare Borgia

Only his omnivorous "appetite for life" can explain Leonardo's decision, in the summer of the following year (1502), to leave Florence and enter the service of Cesare Borgia as "senior military architect and general engineer." Borgia, the notorious son of Pope Alexander VI, had, as commander in chief of the papal army, sought with unexampled ruthlessness to gain control of the Papal States of Romagna and the Marches. Now he was at the peak of his power and, at 27, was undoubtedly the most compelling and at the same time most feared person of his time. Leonardo, twice his age, must have been fascinated by his personality. For 10 months he travelled across the condottiere's territories and surveyed them. In the course of his activity Leonardo sketched some of the city plans and topographical maps that laid the groundwork for modern cartography. At the court of Cesare Borgia, Leonardo also met Niccolò Machiavelli, temporarily stationed there as a political observer for the city of Florence.

In the spring of 1503 Leonardo returned to Florence to make an expert survey of a project for diverting the Arno River behind Pisa so that the city, then under siege by the Florentines, would be deprived of access to the sea. The plan proved unworkable, but Leonardo's activity led him to a much more significant theme, one that served peace rather than war; the project, first advanced in the 13th century and now again under consideration, was to build a large canal that would bypass the unnavigable stretch of the Arno and connect Florence by water with the sea. Leonardo developed his ideas in a series of studies; with panoramic views of the river bank, which are also landscape sketches of great artistic charm, and with exact measurements of the terrain, he produced a map in which the route of the canal (with its transit through the mountain pass of Serravalle) was shown. The project, considered time and again in subsequent centuries, was never carried out, but centuries later the express highway from Florence to the sea was built over the exact route Leonardo chose for his canal.

That same year (1503), however, Leonardo also received a prized commission: to paint a mural for the Hall of the Five Hundred in Florence's Palazzo Vecchio; a historical scene of monumental proportions (at 23 × 56 feet [7 × 17 metres], it would have been twice as large as the "Last Supper"). For three years he worked on this "Battle of Anghiari"; like its intended complementary painting, Michelangelo's "Battle of Cascina," it remained unfinished. But the cartoon and the copies showing the main scene of the battle, the fight for the standard, were for a long time, to quote the sculptor Benvenuto Cellini, "the school of the world." These same years saw the portrait of "Mona Lisa" and a painting of a standing "Leda," which was not completed and has survived only in copies.

Scientific study during the Florentine period

The Florentine period was also, however, a time of intensive scientific study; Leonardo did dissections in the hospital of Sta. Maria Nuova and broadened his anatomical work into a comprehensive study of the structure and function of the human organism. He made systematic observations of the flight of birds, concerning which he planned a treatise. Even his hydrological studies, "on the nature and movement of water," broadened into research on the physical properties of water, especially the laws of currents, which he compared with those pertaining to air. These were also set down in his own collection of data, contained in the so-called Leicester Codex in Holkham Hall, Norfolk, England.

Second Milanese period (1506–13). Thus, during these years in Florence, Leonardo's productivity was also marked by his "creative dualism." Only sporadically did he work at his paintings. When, in May 1506, Charles d'Amboise, governor of the King of France in Milan, asked and was granted permission by the Signoria in Florence for Leonardo to go for a time to Milan, the artist had no hesitation about accepting the invitation. But what was originally a limited period of time became a permanent move under the stress of political circumstances. Florence let Leonardo go, and the monumental "Battle of Anghiari" remained unfinished. Unsuccessful technical experiments with paints seem to have impelled Leonardo to stop working on the mural. One cannot otherwise explain his abandonment of this great work—great both in conception and in realization.

Leonardo spent six years in Milan, interrupted only by a six-month stay in Florence in the winter of 1507–08, where he helped the sculptor Giovanni Francesco Rustici execute his bronze statues for the Florence Baptistery but did not resume work on the "Battle of Anghiari." Honoured and admired by his patrons Charles d'Amboise and King Louis XII, who gave him a yearly stipend of 400 ducats, Leonardo never found his duties onerous. They were limited to advice in architectural matters, tangible evidence of which are plans for a palace-villa for Charles d'Amboise and perhaps also sketches for an oratory for the church of Sta. Maria alla Fontana, which Charles funded. Leonardo also looked into an old project revived by the French governor: the Adda canal that would link Milan with Lake Como by water.

In Milan he did very little as a painter: two Madonnas, which he promised the King of France, were never painted. He continued to work on the paintings of the "Virgin and Child with St. Anne" and "Leda," which he had brought with him from Florence, as copies from the Lombard school of that period attest. Again Leonardo gathered pupils around him. With Ambrogio de Predis he completed a second version of "The Virgin of the Rocks" (1508), in the course of which protracted litigation between the purchasers and the artists had a happy ending. Of his older disciples, Bernardino de' Conti and Salai were again in his studio; new pupils came, among them Cesare da Sesto, Giampetrino, Bernardino Luini, and the young nobleman Francesco Melzi, Leonardo's most faithful friend and companion until his death.

The tomb sculpture for Trivulzio

An important commission in sculpture came his way. Gian Giacomo Trivulzio had returned victoriously to Milan as marshal of the French army and a bitter foe of Ludovico Sforza. He commissioned Leonardo to sculpt his tomb, which was to take the form of an equestrian statue and be placed in the mortuary chapel donated by Trivulzio to the church of S. Nazaro Maggiore. But after years of preparatory work on the monument, for which a number of significant sketches have survived, the Marshal himself gave up the plan in favour of a more modest one; so this undertaking, too, remained unfinished. Leonardo must have felt keenly this second disappointment in his work as a sculptor.

Compared with his almost cursory work in art, Leonardo's scientific activity flourished. His studies in anatomy achieved a new dimension in his collaboration with a famous anatomist from Pavia, Marcantonio della Torre. He outlined a plan for an overall work that would include not only exact, detailed reproductions of the human body and its organs but would also include comparative anatomy and the whole field of physiology. He even thought he would finish his anatomical manuscript in the winter of 1510–11. Beyond that, his manuscripts are replete with mathematical, optical, mechanical, geological, and botanical studies that must be understood as data for his "perceptual cosmology." This became increasingly actuated by a central idea: the conviction that force and motion as basic mechanical functions produce all outward forms in organic and inorganic nature and give them their shape and, furthermore, the recognition that these functioning forces operate in accordance with orderly, harmonious laws.

Last years (1513–19). In 1513 political events—the temporary ouster of the French from Milan—caused the now 60-year-old Leonardo to move again. At the end of the year he went to Rome, accompanied by his pupils Melzi and Salai as well as by two studio assistants, hoping to find employment there through his patron, Giuliano de' Medici, brother of the new pope Leo X. Giuliano gave him a suite of rooms in his residence, the Belvedere, in the Vatican. He also gave him a considerable monthly stipend, but no large commissions came to him. For three years Leonardo remained in the Eternal City, off to one side, while Donato Bramante was building St. Peter's, Raphael was painting the last rooms of the Pope's new apartments, Michelangelo was struggling to complete the tomb of Pope Julius, and many younger artists such as Peruzzi, Timoteo Viti, and Sodoma were active there. Drafts of embittered letters betray the disappointment of the aging master who worked in his studio on mathematical studies and technical experiments or, strolling through the city, surveyed ancient monuments. A magnificently executed map of the Pontine Marshes (Royal Library, Windsor Castle; 12684) suggests that Leonardo was at least a consultant for a reclamation project that Giuliano de' Medici ordered in 1514. On the other hand, there were sketches for a spacious residence for the Medici in Florence, who had returned to power there in 1512. But this did not go beyond the stage of preliminary sketches and never came to pass. Leonardo seems to have resumed his friendship with Bramante, but the latter died in 1514. And there is no record of Leonardo's relations with any other artists in Rome.

Service with King Francis I of France

In a life of such loneliness, it is easy to understand why Leonardo, despite his 65 years, decided to accept the invitation of the young king Francis I to enter his service in France. At the end of 1516 he left Italy forever, together with his most devoted pupil, Francesco Melzi. Leonardo spent the last three years of his life in the small residence of Cloux (later called Clos-Lucé), near the King's summer palace at Amboise on the Loire. *Premier peintre, architecte et méchanicien du Roi* ("first painter, architect, and mechanic of the King") was the proud title he bore; yet the admiring King left him complete freedom of action. He did no more painting or at most completed the painting of the enigmatic, mystical "St. John the Baptist," which the Cardinal of Aragon, when he visited Amboise, saw in Leonardo's studio along with the "Mona Lisa" and the "Virgin and Child with St. Anne."

For the King he drew up plans for the palace and garden of Romorantin, destined to be the widow's residence of the Queen Mother. But the carefully worked-out project, combining the best features of Italian-French traditions in palace and landscape architecture, had to be halted because the region was threatened with malaria.

Leonardo still made sketches for court festivals, but the King treated him in every respect as an honoured guest. Decades later, Francis I talked with the sculptor Benvenuto Cellini about Leonardo in terms of the utmost admiration and esteem. Leonardo spent most of his time arranging and editing his scientific studies. The final drafts for his treatise on painting and a few pages of the anatomy appeared. Consummate drawings such as the "Floating Figure" (Royal Library, Windsor Castle; 12581) are the final testimonials to his undiminished genius. In the so-called "Visions of the End of the World," or "Deluge" (Royal Library, Windsor Castle), he depicts with overpowering pictorial imagination the primal forces that rule nature.

On May 2, 1519, Leonardo died at Cloux. He was laid to rest in the palace church of Saint-Florentin. But the church was devastated during the French Revolution and completely torn down at the beginning of the 19th century. Hence, his grave can no longer be located. Francesco Melzi fell heir to his artistic and scientific estate.

ANALYSIS AND EVALUATION OF LEONARDO'S ACHIEVEMENT

Painting. Leonardo's total output in painting is really not large; only 17 of the paintings that have survived can be definitely attributed to him, and several of them are unfinished. Two of his most important works—the "Battle of Anghiari" and the "Leda," neither of them completed—have only survived in copies. Yet these few creations have established the unique fame of a man whom Vasari, in his *Lives,* dividing art history into three ages, placed in the last "golden age of the arts." His works, unaffected by all the vicissitudes of aesthetic doctrines in subsequent centuries, have stood out in all periods and all countries as consummate masterpieces of painting.

Leonardo's art of expression

The many testimonials to Leonardo, ranging from Vasari to Peter Paul Rubens, Johann Wolfgang von Goethe, and Eugène Delacroix, make it unmistakably clear that it has been, above all, Leonardo's art of expression that has called forth the utmost admiration. It is, in fact, the core of his formation as a painter—from his earliest beginnings to his last work. This expression was nurtured by his power of invention but also by every technical means: drawing, colour, use of light and shadow. To Leonardo, expression became a key concept of art; it also included the basic demands of truth, beauty, and accuracy in everything depicted.

What Leonardo was striving for was already revealed in his angel in Verrocchio's "Baptism of Christ" (*c.* 1474–75): in the natural structuring of the angel's body based on movement in several directions, in the relaxation of his attitude, and in his glance, which takes in what is occurring but at the same time is directed inward. In his landscape segment in the same picture, Leonardo also found a new expression for "nature experienced," in reproducing the forms he perceived as if through a veil of mist. The landscape study (Uffizi, Florence) dated 1473, a pen drawing, foreshadows in its treatment of transparent atmosphere by a 21-year-old his telling ability to transform perceived phenomena into convincing graphic forms.

In the "Madonna Benois" (1478) Leonardo succeeded in giving an old traditional type of picture a new, unusually charming, and expressive mood by showing the child Jesus reaching for the flower in Mary's hand in a sweet and tender manner.

His "Portrait of Ginevra de' Benci" (*c.* 1475–78) opened new paths for portrait painting with his singular linking of nearness and distance.

The emaciated body of his "St. Jerome" (*c.* 1480) is presented with realistic truth based on his sober and objective studies in anatomy; gesture and look give Jerome an unrivalled expression of transfigured sorrow.

The interplay of mimicry and gesture—"physical and spiritual motion," in Leonardo's words—is also the chief concern of his first large creation containing many figures, "The Adoration of the Magi" (1481). Never finished, the painting nevertheless affords rich insight into the master's subtle methods of work. The various aspects of the scene are "built up" from the base with very delicate, paper-thin layers of paint in chiaroscuro (the balance of light and shadow) relief. The main treatment of the Virgin and Child group and the secondary treatment of the surrounding groups are clearly set apart with a masterful sense of composition; yet thematically they are closely interconnected: the bearing and expression of the figures—most striking in the group of praying shepherds—depict all degrees and levels of profound amazement. "The Virgin of the Rocks" in its first version in the Louvre is the work that reveals Leonardo's painting art at its purest. The painting, according to Leonardo's first contract with the Confraternity of the Immaculate Conception, was to be the central panel of a large work for their chapel in the church of S. Francesco and was done in the years *c.* 1483–85. It never arrived, however, at the place it was originally destined for. It seems to have been prematurely taken from the Confraternity, perhaps by some highly placed interested party who removed it from Leonardo's workshop. Instead of this first painting, Leonardo and Ambrogio de Predis painted a second, slightly revised version, probably begun around 1494. This one gave rise to a 10-year litigation between the artist and the Confraternity regarding the price, a dispute that was not settled until 1506 in favour of Leonardo; whereupon, two years later, the painting was delivered as per contract. This second version remained in the chapel of S. Francesco until the Confraternity was dissolved (1781), and then, after changing owners frequently, it came finally in 1880 to the National Gallery in London.

The two versions of "The Virgin of the Rocks"

"The Virgin of the Rocks" depicts the apocryphal legend of the meeting in the wilderness between the boy John and the equally young Jesus returning home from Egypt. Leonardo's artistry makes of this theme a vision that the true believer experiences when he contemplates the devotional picture. In the visionary character of the picture lies the secret of its effect: it presents not a "reality" but a "manifestation." Leonardo uses every artistic means at his disposal to emphasize the visionary nature of the scene. The soft colour tones (his famous sfumato), the dim light of the cave from which the figures emerge bathed in light, their quiet attitude, the meaningful gesture with which the angel (the only one facing the viewer) points to John as the intercessor between the Son of God and humanity—all this combines, in a patterned and formal way, to achieve an effect of the highest expressiveness.

The "Last Supper"

Leonardo's "Last Supper" is among the most famous paintings in the world. In its monumental simplicity, the composition of the scene is masterful; the power of its effect comes from the striking contrast in the attitudes of the 12 disciples as counterposed to Christ. Leonardo did not choose the portrayal of the traitor Judas customary in the iconographic tradition; he portrayed, rather, that moment of highest tension as related in the New Testament, "One of you which eateth with me will betray me." All of the Apostles—as human beings who do not understand what is about to occur—are agitated, whereas Christ alone, conscious of his divine mission, sits in lonely, transfigured serenity. Only one other being shares the secret knowledge: Judas, who is both part of and yet excluded from the movement of his companions; in this isolation he becomes the second lonely figure—the guilty one—of the company.

In the profound conception of his theme, in the perfect yet seemingly simple arrangement of the individuals, in the temperaments of the Apostles highlighted by gesture and mimicry, in the drama and at the same time the sublimity of the treatment, Leonardo attained a height of expression that has remained a model of its kind. Untold painters in succeeding generations, among them great masters such as Rubens and Rembrandt, marvelled at Leonardo's composition and were influenced by it. The painting also inspired some of Goethe's finest pages of descriptive prose. It has become widely known through countless reproductions and prints, the most important being those produced by Raffaello Morghen in 1800. Thus, the "Last Supper" has become part of humanity's common heritage and remains today one of the world's outstanding paintings.

Technical deficiencies in the execution of the work have not lessened its fame. Leonardo was uncertain about the technique he should use. He bypassed fresco painting, which, because it is executed on fresh plaster, demands quick and uninterrupted painting, in favour of another technique he had developed: tempera on a base mixed by himself on the stone wall. This procedure proved unsuccessful, inasmuch as the base soon began to be loosened from the wall. Damage appeared by the beginning of the 16th century, and deterioration soon set in. By the middle of the century the work was called a ruin. Later, inadequate attempts at restoration only aggravated the situation, and not until the most modern restoration techniques were applied after World War II was the process of decay halted.

In the Florence years between 1500 and 1506, four great creations appeared that confirmed and heightened Leonardo's fame: the "Virgin and Child with St. Anne" (Louvre), "Mona Lisa," "Battle of Anghiari," and "Leda." Even before it was completed, the "Virgin and Child with St. Anne" won the critical acclaim of the Florentines; the monumental plasticity of the group and the calculated effects of dynamism and tension in the composition made it a model that inspired Classicists and Mannerists in equal measure. The "Mona Lisa" became the ideal type of portrait, in which the features and symbolic overtones of the person painted achieved a complete synthesis. The young Raphael sketched the work in progress, and it served as a model for his "Portrait of Maddalena Doni." Similarly, the "Leda" became a model of the *figura serpentinata* ("sinuous figure")—that is, a figure built up from several intertwining views. It influenced such classical artists as Raphael, who drew it, but it had an equally strong effect on Mannerists such as Jacopo Pontormo.

The "Mona Lisa"

In the "Battle of Anghiari" (1503–06) Leonardo's art of expression reached its high point. The preliminary drawings—many of which have been preserved—reveal Leonardo's lofty conception of the "science of painting"; the laws of equilibrium that he had probed in his studies in mechanics were put to artistic use in this painting. The "centre of gravity" lies in the group of flags fought for by all the horsemen. For a moment the intense and expanding movement of the swirl of riders seems frozen; this passing moment, the transition from one active movement to the next, is uniquely interpreted.

On the other hand, Leonardo's studies in anatomy and physiology influenced his representation of human and animal bodies, particularly when they were in a state of excitement. He studied and described extensively the baring of teeth and puffing of lips as signs of animal and human anger. On the painted canvas, rider and horse, their features distorted, are remarkably similar in expression.

The highly imaginative trappings take the event out of the sphere of the historical into a timeless realm. Thus, the "Battle of Anghiari" became the standard model for a cavalry battle. Its composition has influenced many painters: from Rubens in the 17th century, who made the most impressive copy of the scene from Leonardo's now-lost cartoon, to Delacroix in the 19th century.

After 1507—in Milan, Rome, and France—Leonardo did very little painting. He did resume work on the Leda theme during his years in Milan and sketched a variation, the "Kneeling Leda." The drawings he prepared—revealing examples of his late style—have a curious, enigmatic sensuality. Perhaps in Rome he began the "St. John the Baptist," which he completed in France. Bursting all the boundaries of usual painting tradition, he presented Christ's forerunner as the herald of a mystic oracle; his was an "art of expression" that seemed to strive consciously to bring out the hidden ambiguity of the theme.

The last manifestation of Leonardo's art of expression was in his "Visions of the End of the World," a series

Last manifestation of Leonardo's art of expression

of pictorial sketches that took the end of the world as its theme. Here Leonardo's power of imagination—born of reason and fantasy—attained its highest level. The immaterial forces in the cosmos, invisible in themselves, appear in the material things they set in motion. What Leonardo had observed in the swirling of water and eddying of air, in the shape of a mountain boulder and in the growth of plants now assumed gigantic shape in cloud formations and rainstorms. The framework of the world splits asunder, but even its destruction occurs—as the monstrously "beautiful" forms of the unleashed elements show—in accordance with the self-same laws of order, harmony, and proportion that presided at its creation and that govern the life and death of every created thing in nature. Without any model, these "visions" are the last and most original expressions of Leonardo's art—an art in which his perception based on *saper vedere* seems to have come to fruition.

Sculpture. That Leonardo worked as a sculptor from his youth on is borne out by his own statements and those of other sources. In the introduction to his *Treatise on Painting* he gives painting precedence over sculpture in the hierarchy of the arts; yet he emphasizes that he practices both arts equally. A small group of generals' heads in marble and plaster, works of Verrocchio's followers, are sometimes linked with Leonardo because a lovely drawing on the same theme from his hand suggests such a connection. But the inferior quality of this group rules out an attribution to the master. Not a trace has remained of the heads of women and children that, according to Vasari, Leonardo modelled in clay in his youth.

The two great sculptural projects to which Leonardo devoted himself wholeheartedly stood under an unlucky star; neither the huge, bronze equestrian statue for Francesco Sforza, on which he worked until 1494, nor the monument for Marshal Trivulzio, on which he was busy in the years 1506–11, were brought to completion. Leonardo kept a detailed diary about his work on the Sforza horse; it came to light with the rediscovery of the Madrid MS. 8936. Text and drawings both show Leonardo's wide experience in the technique of bronze casting but at the same time reveal the almost utopian nature of the project. He wanted to cast the horse in a single piece, but the gigantic dimensions of the steed presented insurmountable technical problems. Indeed, Leonardo remained uncertain of the problem's solution to the very end.

The greatness of Leonardo's concept of sculpture

The drawings for these two monuments reveal the greatness of Leonardo's concept of sculpture. Exact studies of the anatomy, movement, and proportions of a live horse—Leonardo even seems to have thought of writing a treatise on the horse—preceded the sketches for the monuments. Leonardo pondered the merits of two types, the galloping or trotting horse, and in both cases decided in favour of the latter. These sketches, superior in the suppressed tension of horse and rider to the achievements of Donatello's Gattamelata and Verrocchio's Colleoni sculptures, are among the most beautiful and significant examples of Leonardo's art. Unquestionably—as ideas—they exerted a very strong influence on the development of equestrian statues in the 16th century.

A small bronze of a galloping horseman in Budapest is so close to Leonardo's style that, if not from his own hand, it must have been done under his immediate influence (perhaps by Giovanni Francesco Rustici). Rustici, according to Vasari, was Leonardo's zealous student and enjoyed his master's help in sculpting his large group in bronze of "St. John the Baptist Teaching" over the north door of the Baptistery in Florence. There are, indeed, discernible traces of Leonardo's influence in John's stance, with the unusual gesture of his upward pointing hand, and in the figure of the bald-headed Levite. Moreover, an echo of Leonardo's inspiration is unmistakable in the much-discussed and much-reviled wax bust of "Flora" in Berlin. It may have been made in France, perhaps in the circle of Rustici, who entered Francis I's service in 1528.

Architecture. Leonardo, who in a letter to Ludovico Sforza applying for service described himself as an experienced architect, military engineer, and hydraulic engineer, was concerned with architectural matters all his life. But his effectiveness was essentially limited to the role of an adviser. Only once—in the competition for the cupola of the Milan cathedral (1487–90)—did he actually consider personal participation; but he gave up this idea when the model he had submitted was returned to him. In other instances, his claim to being a practicing architect involved sketches for representative secular buildings: for the palace of a Milanese nobleman (around 1490), for the villa of the French governor in Milan (1507–08), and for the Medici residence in Florence (1515). Finally, there was his big project for the palace and garden of Romorantin in France (1517–19). Especially in this last named, Leonardo's pencil sketches clearly reveal his mastery of technical as well as artistic architectural problems; the view in perspective (at Windsor Castle) gives an idea of the magnificence of the site.

Activity as a military engineer

Leonardo was also quite active as a military engineer, beginning with the years of his stay in Milan. But no definite examples of his work can be adduced. Not until the discovery of the Madrid notebooks was it known that in 1504, sent probably by the Florence governing council, he stood at the side of the Lord of Piombino when the city's fortifications system was repaired and that Leonardo suggested a detailed plan for overhauling it. Finally, his studies for large-scale canal projects in the Arno region and in Lombardy show that he was also an expert in hydraulic engineering.

But what really characterizes Leonardo's architectural studies and makes them stand out is their comprehensiveness; they range far afield and embrace every type of building problem of his time. Furthermore, there frequently appears evidence of Leonardo's impulse to teach: he wanted to collect his writings on this theme in a theory of architecture. This treatise on architecture—the initial lines of which are in MS. B (Institut de France, Paris), a model book of the types of sacred and profane buildings—was to deal with the entire field of architecture as well as with the theory of forms and construction and was to include such items as urbanism, sacred and profane building, and a compendium of the important individual elements (for example, domes, steps, portals, and windows).

In the fullness and richness of their ideas, Leonardo's architectural studies offer an unusually wide-ranging insight into the architectural achievements of his epoch. Like a seismograph, his observations sensitively register all themes and problems. For almost 20 years he was associated with Bramante at the court of Milan and again met him in Rome in 1513–14; he was closely associated with such other distinguished architects as Francesco di Giorgio, Giuliano da Sangallo, Giovanni Antonio Amadeo, and Luca Fancelli. Thus, he was brought in closest touch with all of the most significant building undertakings of the time. Since Leonardo's architectural drawings extend over his whole life, they span precisely that developmentally crucial period—from the 1480s to the second decade of the 16th century—in which the principles of the classical style were formulated and came to maturity. That this genetic process can be followed in the ideas of one of the greatest men of the period lends Leonardo's studies their distinctive artistic value and their outstanding historical significance.

Science. *Science of painting.* Notwithstanding Leonardo's abundant scientific activity, one must never lose sight of the fact that it was the intellectual output of a man who proudly and consciously felt himself an artist throughout his life. And he described himself as such. He first came in contact with science as an artist, in the task he set himself of writing a treatise on painting.

Treatise on Painting

Leonardo's famous book on painting, in the form known and read today, is not an original work by the master but a compilation of texts from various manuscripts by Leonardo, collected and arranged with loving care by his disciple and heir, Francesco Melzi. It is the Codex Urbinas Latinus 1270, now in the Vatican Library. It was prepared around 1540–50, but from its form one can see that it was still an unfinished rather than a completed manuscript. Many original texts known to exist are missing; whole sections of Leonardo's overall plan are not included.

The first printed edition of the treatise in Melzi's version, omitting the long introductory chapter concerning the "pecking order" among the arts, appeared in a luxurious binding in 1651 in Paris, published by Raffaelo du Fresne with illustrations after drawings by Nicolas Poussin. The first complete edition of Melzi's text did not appear until 1817, published by Guglielmo Manzi in Rome. The two standard modern editions are that of Emil Ludwig, three volumes, Vienna, 1882 (with German translation); and that of A. Philip McMahon, Princeton, 1956, two volumes (facsimile of the Codex Urbinas and English translation).

Leonardo's plan envisaged a much broader treatment of the theme, as his own allusions to it indicate. For, in addition to detailed practical instructions for painting and drawing, the treatise was to deal with every area involving the artist's perception and experience, which he could then convey as acquired criteria. Three main problems form the keynote of the work: the definition of painting as a science, which is briefly outlined above; the theory of the mathematical basis of painting—that is, geometry, perspective, and optics—with the systematic study of light and shadow, colour, and aerial perspective; and the theory of forms and functions in organic and inorganic nature, as they are explained and made comprehensible to the painter trained in *saper vedere.* This theory of the forms and functions of the visible world sought first of all to describe the animal world, including man; next it sought to include the plant world; finally it endeavoured to explain how such phenomena of inorganic nature as water and earth, air and fire came into being.

Drawings for the *Treatise*

In the drawings for the *Treatise on Painting,* extending from the earliest Milan period to the final years of Leonardo's life in France, the progressive broadening and deepening of the theme can be followed. Many drawings were placed by the side of the text, and some of them were coloured; many studies of nature that are admired as art works, such as the famous rain landscape (Windsor Castle; 12409) or the "Foliage" (Royal Library, Windsor Castle; 12431), can be identified as illustrations for the treatise. Manuscript C in the Institut de France, Paris, with its diagrams of the blending of lights and shadows, likewise represents a segment of this textbook. Leonardo's so-called grotesque heads are also closely linked with the treatise. They have often been erroneously described as caricatures; but actually, for the most part, they represent types and only occasionally individuals. They are variations of the human face in its gradations between the poles of the beautiful and ugly, the normal and abnormal, the dignified and vulgar. They are also related to anatomical-physiological studies, in which old age—with wrinkled skin and bulging tendons—is contrasted with youth. Representation of the human being was to be treated at length: his body, his proportions, his organs and their functions but also his attitudes in physical and spiritual movement. Here Leonardo's artistic and scientific aims intertwine.

Anatomical studies and drawing. Leonardo's anatomical studies are perhaps the best way of revealing the process by which, in Leonardo's mind, an increasing differentiation set in among his diverse spheres of interest; but it was a differentiation in which the seemingly divergent areas of study—likewise on a higher level—always remained interrelated. Thus, Leonardo's study of anatomy, originally pursued for his training as an artist, quickly grew into an independent area of research. As his sharp eye uncovered the structure of the human body, Leonardo became fascinated by the *figura istrumentale dell' omo* ("man's instrumental figure"), and he sought to probe it and present it as a creation of nature. The early studies dealt chiefly with the skeleton and muscles; yet even at the outset Leonardo combined anatomical with physiological researches. From observing the static structure, Leonardo proceeded to study the functions exercised by the individual parts of the body as they bring into play the organism's mechanical activity. This led him finally to the study of the internal organs; among them he probed most deeply into the brain, heart, and lungs as the "motors" of the senses and of life. He did practical work in anatomy on the dissection table in Milan, then in the hospital of Sta. Maria Nuova in Florence, and again in Milan and Pavia, where he received counsel and inspiration from the physician-anatomist Marcantonio della Torre. By his own admission he dissected 30 corpses in his lifetime, thus acquiring an astonishing range of experience on his own. This experience was distilled in the famous anatomical drawings, which are among the most significant achievements of Renaissance science. These drawings, among his *dimostrazione,* are based on a curious connection between natural and abstract representation; sections in perspective, reproduction of muscles as "strings" or the indication of hidden parts by dotted lines, and finally a specifically devised hatching system enable him to represent any part of the body in transparent layers that afford an "insight" into the organ. Here Leonardo's mastery of drawing proved most useful. The genuine value of these *dimostrazione* and their superiority to descriptive words—as Leonardo proudly emphasized—lay in the fact that they were able to synthesize a multiplicity of individual experiences at the dissecting table and make the data immediately and accurately visible. The effect is unlike that of all dead anatomical preparations; in this way the "live quality" of the organism is retained.

Value of the *dimostrazione*

This great picture chart of the human body was what Leonardo envisaged as a *cosmografia del minor mondo* ("cosmography of the microcosm"). From the advanced portions that have survived, it is apparent how much and how long it occupied his mind. And it provided the basic principles for modern scientific illustration. Leonardo has not sufficiently received his due in this domain. Thanks to a method of seeing that was peculiarly his own, he elevated the art of drawing into a means of scientific investigation and teaching of the highest quality.

Mechanics and cosmology. With Leonardo, mechanics also proceeds from artistic practice, with which he became quite familiar as an architect and engineer. Throughout his life Leonardo was an inventive builder; he was thoroughly at home in the principles of mechanics of his epoch and contributed in many ways to advancing them.

His model book on the elementary theory of mechanics, which appeared in Milan at the end of the 1490s, was discovered in the Madrid Codex 8937. Its importance lay less in its description of specific machines or work tools than in its use of demonstration models to explain the basic mechanical principles and functions employed in building machinery. Leonardo was especially concerned with problems of friction and resistance. These elements—screw threads, gears, hydraulic jacks, swivelling devices, transmission gears, and the like—are described individually or in various combinations; and here, too, drawing takes precedence over the written word. As in his anatomical drawings, Leonardo develops definite principles of graphic representation—stylization, patterns, and diagrams—that guarantee a precise demonstration of the object in question.

Importance of primal mechanical forces to his thought

In the course of years his interest in pure mechanics merged increasingly with an interest in applied mechanics. Leonardo realized that the mechanical forces at work in the basic laws of mechanics operate everywhere in the organic and inorganic world. They determine animate and inanimate nature alike as well as man. Leonardo wrote on a page of his treatise on anatomy:

> See to it that the book of the principles of mechanics precedes the book of force and movement of man and the other living creatures, for only in that way will you be able to prove your statements.

So, finally, "force" became the key concept for Leonardo; as *virtù spirituale* ("spiritual property"), it shaped and ruled the cosmos.

Wherever Leonardo probed the phenomena of nature, he recognized the existence of primal mechanical forces that govern the shape and function of the universe: in his studies on the flight of birds, in which his youthful idea of the feasibility of a flying apparatus took shape and led to exhaustive research into the element of air; in his studies of water, the *vetturale della natura* ("conveyor of nature"), in which he was as much concerned with the physical properties of water as with its laws of motion and currents; in his researches on the laws of growth of plants and trees as well as the geological structure of earth and hill formations; and finally in his observation of air currents, which evoked the image of the flame of a candle or the picture of a wisp of cloud and smoke. In his drawings, especially in his studies of whirlpools, based on numerous experiments he undertook, Leonardo again found a stylized form of representation that was uniquely his own: this involved breaking down a phenomenon into its component parts—the traces of water or eddies of the whirlpool—yet at the same time preserving the total picture, analytic and synthetic vision.

Thus, for all the separate individual realms of his knowledge, Leonardo's science offered a unified picture of the world: a cosmogony based on *saper vedere*. Its final wisdom is that all the workings of nature are subject to a law of necessity and a law of order that the *Primo Motore*, the divine "Prime Mover," created. "Marvelous is Thy justice, O Prime Mover! Thou hast seen to it that no power lacks the order and value of your necessary governance."

Leonardo as artist-scientist. As the 15th century expired, Scholastic doctrines were in decline, and Humanistic scholarship was on the rise. Leonardo, however, was part of an intellectual circle that developed a third, specifically modern form of cognition. In his view the artist—as transmitter of the true and accurate data of experience acquired by visual observation—played a significant part. With this sense of the artist's high calling, Leonardo approached the vast realm of nature to probe its secrets. His utopian idea of transmitting in encyclopaedic form the knowledge thus won was still bound up with medieval Scholastic conceptions, but the results of his research were among the first great achievements of the thinking of the new age because they were based on the principle of experience in an absolutely new way and to an unprecedented degree.

Finally, Leonardo, although he made strenuous efforts to teach himself and become erudite in languages, natural science, mathematics, philosophy, and history, as a mere listing of the wide-ranging contents of his library demonstrates, remained an empiricist of visual observation. But precisely here—thanks to his genius—he developed his own "theory of knowledge," unique in its kind, in which art and science form a synthesis. In the face of the overall achievements of Leonardo's creative genius, the question of how much he finished or did not finish becomes pointless. The crux of the matter is his intellectual force—self-contained and inherent in every one of his creations. This force has remained constantly operative to the present day. (L.H.H.)

MAJOR WORKS

Paintings. "The Annunciation" (*c.* 1472–77; Uffizi, Florence); "The Annunciation" (*c.* 1472–77; Louvre, Paris); "Madonna with the Carnation" (*c.* 1474; Alte Pinakothek, Munich); "Portrait of Ginevra de' Benci" (*c.* 1475–78; National Gallery of Art, Washington, D.C.); "Madonna Benois" (1478–after 1500; Hermitage, St. Petersburg); "St. Jerome" (*c.* 1480; Vatican Museums, Rome); "The Adoration of the Magi" (1481; Uffizi); "The Virgin of the Rocks" (*c.* 1483–85; Louvre); "The Musician" (*c.* 1490; Pinacoteca Ambrosiana, Milan); "Lady with an Ermine" ("Cecilia Gallerani"; *c.* 1490; Muzeum Narodowe, Kraków, Poland); "The Virgin of the Rocks" (1494–1508; National Gallery, London); "Last Supper" (1495–97; Sta. Maria delle Grazie, Milan); decoration of the Sala delle Asse (1498; Castello Sforzesco, Milan); "The Virgin and Child with St. Anne" (cartoon, *c.* 1499; National Gallery); "Virgin and Child with St. Anne" (*c.* 1501–12; Louvre); "Mona Lisa" ("La Gioconda"; 1503–06; Louvre); "St. John the Baptist" (before 1517; Louvre). Lost: "Madonna with the Yarn-Winder" (1501; best copy in the Duke of Buccleuch Collection, Boughton, Kettering); "Leda" (1503–06; best copy at Galleria Borghese, Rome); "Battle of Anghiari" (1503–06; copy at Palazzo Vecchio, Florence).

Drawings and notebooks. Main collections: Institut de France, Paris; British Museum; Uffizi, Florence; Biblioteca Ambrosiana, Milan; Accademia, Venice; Royal Library, Windsor Castle; Biblioteca Reale, Turin; Biblioteca Nacional, Madrid; Victoria and Albert Museum, London.

BIBLIOGRAPHY. ANGELA OTTINO DELLA CHIESA (ed.), *The Complete Paintings of Leonardo da Vinci* (1967, reissued 1985; originally published in Italian, 1967), catalogs the paintings. The standard publication on the drawings is KENNETH CLARK, *The Drawings of Leonardo da Vinci in the Collection of Her Majesty the Queen at Windsor Castle,* 2nd ed., rev. with CARLO PEDRETTI, 3 vol. (1968). A.E. POPHAM (ed.), *The Drawings of Leonardo da Vinci,* 2nd ed. (1947, reissued 1973), is important for the study of Leonardo as a draftsman. A. MARIONI, "I manoscritti di Leonardo da Vinci," in COMITATO NAZIONALE PER LE ONORANZE A LEONARDO DA VINCI NEL QUINTO CENTENARIO DELLA NASCITA, *Leonardo: Saggi e richerche* (1954), is a concise summary of all manuscripts, their facsimile editions, and their chronology and contains other excellent essays by various authors on Leonardo as artist and scientist. *The Madrid Codices,* 5 vol. (1974), contains facsimiles of the codices (vol. 1–2), commentary by LADISLAO RETI (vol. 3), and Reti's transcription and translation of the codices into English (vol. 4–5). Also of interest is LADISLAO RETI, "The Two Unpublished Manuscripts of Leonardo da Vinci in the Biblioteca Nacional of Madrid," *Burlington Magazine,* 110:10–24 (January/February 1968), with extended discussion in the February 1969 issue, pp. 111–191. A. PHILIP MCMAHON (trans.), *Treatise on Painting,* 2 vol. (1956), is a facsimile edition of *Codex Urbinas Latinus 1270,* accompanied by an English translation. KENNETH D. KEELE and CARLO PEDRETTI, *Leonardo da Vinci: Corpus of the Anatomical Studies in the Collection of Her Majesty the Queen at Windsor Castle,* 3 vol. (1978–80), includes a volume of facsimile plates. The best anthologies of Leonardo's literary heritage are EDWARD MCCURDY (ed. and trans.), *The Notebooks of Leonardo da Vinci,* 2nd ed., 2 vol. (1955, reissued 1977); and JEAN PAUL RICHTER (compiler and ed.), *The Literary Works of Leonardo da Vinci,* trans. from Italian, 3rd ed., 2 vol. (1970). A selection of stories, with drawings, from Leonardo's notebooks is found in EMERY KELEN (ed.), *Fantastic Tales, Strange Animals, Riddles, Jests, and Prophecies of Leonardo da Vinci* (1971). MARTIN KEMP (ed.), *Leonardo on Painting: An Anthology of Writings* (1989), is a readable and organized translated collection of Leonardo's notes on art. ISTITUTO GEOGRAFICO DE AGOSTINI, *Leonardo da Vinci* (1956; originally published in Italian, 1938), an exhibition catalog, contains numerous essays and is a richly illustrated compendium of Leonardo's artistic and scientific activity.

The two standard publications on Leonardo sources are LUCA BELTRAMI (ed.), *Documenti e memorie riguardanti la vita e le opere di Leonardo da Vinci in ordine cronologico* (1919); and GEROLAMO CALVI, *I manoscritti di Leonardo da Vinci, dal punto di vista cronologico, storico e biografico* (1925, reissued 1982). Additional sources of information include ETTORE VERGA, *Bibliografia Vinciana, 1493–1930,* 2 vol. (1931, reprinted 1970); and *Raccolta Vinciana,* fascicle 1–20 (1905–64).

Studies of Leonardo's life and works are found in GABRIEL SÉAILLES, *Léonard de Vinci: l'artiste & le savant,* new, rev. and augmented ed. (1912); WOLDEMAR VON SEIDLITZ, *Leonardo da Vinci,* new ed. edited by KURT ZOEGE VON MANTEUFFEL (1935), in German, accompanied by extensive documentation; LUDWIG H. HEYDENREICH, *Leonardo da Vinci,* 2 vol. (1954; originally published in German, 1953); RICHARD MCLANATHAN, *Images of the Universe: Leonardo da Vinci: The Artist as Scientist* (1966); MORRIS PHILIPSON (ed.), *Leonardo da Vinci: Aspects of the Renaissance Genius* (1966), containing valuable contributions to the historical and psychological aspects of Leonardo; V.P. ZUBOV, *Leonardo da Vinci* (1968; originally published in Russian, 1961); C.D. O'MALLEY (ed.), *Leonardo's Legacy: An International Symposium* (1969), a collection of essays exploring various aspects of Leonardo's works; RITCHIE CALDER, *Leonardo & the Age of the Eye* (1970), with emphasis on his artistic as well as his scientific work; LADISLAO RETI (ed.), *The Unknown Leonardo* (1974, reprinted 1990), 10 essays discussing aspects of Leonardo's personality and creativity made evident in the Madrid Codices; CECIL GOULD, *Leonardo: The Artist and the Non-artist* (1975); ROBERT PAYNE, *Leonardo* (1978), an account of Leonardo's career, with several new interpretations; MARTIN KEMP, *Leonardo da Vinci: The Marvellous Works of Nature and Man* (1981); and KENNETH CLARK, *Leonardo da Vinci,* new ed. rev. by MARTIN KEMP (1988). The most informative account of Leonardo's workshop and pupils is WILHELM SUIDA, *Leonardo und sein Kreis* (1929). Leonardo's architectural ventures are examined in CARLO PEDRETTI, *Leonardo da Vinci: The Royal Palace at Romorantin* (1972), and *Leonardo, Architect* (1985; originally published in Italian, 1978), an in-depth survey, while *Leonardo: A Study in Chronology and Style* (1973, reprinted 1982), focuses on his art. EMANUEL WINTERNITZ, *Leonardo da Vinci as a Musician* (1982), studies Leonardo's interest in and explorations of music and musical instruments. His studies of human anatomy and movement are discussed and supplemented by numerous illustrations in KENNETH D. KEELE, *Leonardo da Vinci's Elements of the Science of Man* (1983). A seldom-explored topic, Leonardo's interest in botany, is treated in WILLIAM A. EMBODEN, *Leonardo da Vinci on Plants and Gardens* (1987). A. RICHARD TURNER, *Inventing Leonardo* (1993), treats his posthumous reputation. (L.H.H./Ed.)

Libraries

Libraries, once known primarily as storehouses for books and periodicals, have changed dramatically since the middle of the 20th century. From their historical beginnings as places to keep the business, legal, historical, and religious records of a civilization, libraries have emerged as a far-reaching body of information resources and services that do not require a building. Rapid developments in computers, telecommunications, and other technologies have made it possible to store and retrieve information in many different forms and from any place with a computer and a telephone connection. The terms *digital library* and *virtual library* have begun to be used to refer to the vast collections of information to which people gain access over the Internet, cable television, or some other type of remote electronic connection.

This article provides a history of libraries from their founding in the ancient world through the latter half of the 20th century, when both technological and political forces radically reshaped library development. It offers an overview of several types of traditional libraries and explains how libraries collect, organize, and make accessible their collections. Further discussion of the application of the theory and technology of information science in libraries and related fields is included in the article INFORMATION PROCESSING AND INFORMATION SYSTEMS.

This article is divided into the following sections:

The changing role of libraries

Libraries are collections of books, manuscripts, journals, and other sources of recorded information. They commonly include reference works, such as encyclopaedias that provide factual information and indexes that help users find information in other sources; creative works, including poetry, novels, short stories, music scores, and photographs; nonfiction, such as biographies, histories, and other factual reports; and periodical publications, including magazines, scholarly journals, and books published as part of a series. As home use of records, CD-ROMs, and audiotapes and videotapes has increased, library collections have begun to include these and other forms of media, too.

Libraries were involved early in exploiting information technologies. For many years libraries have participated in cooperative ventures with other libraries. Different institutions have shared cataloging and information about what each has in its collection. They have used this shared information to facilitate the borrowing and lending of materials among libraries. Librarians have also become expert in finding information from on-line and CD-ROM databases.

As society has begun to value information more highly, the so-called information industry has developed. This industry encompasses publishers, software developers, on-line information services, and other businesses that package and sell information products for a profit. It provides both an opportunity and a challenge to libraries. On the one hand, as more information becomes available in electronic form, libraries no longer have to own an article or a certain piece of statistical information, for example, to obtain it quickly for a user. On the other hand, members of the information industry seem to be offering alternatives to libraries. A student with her own computer can now go directly to an on-line service to locate, order, and receive a copy of an article without ever leaving her home.

Although the development of digital libraries means that people do not have to go to a building for some kinds of information, users still need help to locate the information they want. In a traditional library building, a user has access to a catalog that will help locate a book. In a digital library, a user has access to catalogs to find traditional library materials, but much of the information on, for example, the Internet can not be found through one commonly accepted form of identification. This problem necessitates agreement on standard ways to identify pieces of electronic information (sometimes called meta-data) and the development of codes (such as HTML [Hypertext Markup Language] and SGML [Standard Generalized Markup Language]) that can be inserted into electronic texts.

For many years libraries have bought books and periodicals that people can borrow or photocopy for personal use. Publishers of electronic databases, however, do not usually sell their product, but instead they license it to libraries (or sites) for specific uses. They usually charge libraries a per-user fee or a per-unit fee for the specific amount of information the library uses. When libraries do not own these resources, they have less control over whether older information is saved for future use—another important cultural function of libraries. In the electronic age, questions of copyright, intellectual property rights, and the economics of information have become increasingly important to the future of library service.

Increased availability of electronic information has led libraries, particularly in schools, colleges, and universities, to develop important relationships to their institutions' computer centres. In some places the computer centre is the place responsible for electronic information and the library is responsible for print information. In some educational institutions librarians have assumed responsibility for both the library collection and computer services.

As technology has changed and allowed ever new ways of creating, storing, organizing, and providing information, public expectation of the role of libraries has increased. Libraries have responded by developing more sophisticated on-line catalogs that allow users to find out whether or not a book has been checked out and what other libraries have it. Libraries have also found that users want information faster, they want the full text of a document instead of a citation to it, and they want information that clearly answers their questions. In response, libraries have provided Selective Dissemination of Information (SDI) services, in which librarians choose information that may be of interest to their users and forward it to them before the users request it.

The changes in libraries outlined above originated in the United States and other English-speaking countries. But electronic networks do not have geographic boundaries, and their influence has spread rapidly. With Internet connections in Peking (Beijing), Moscow, and across the globe, people who did not have access to traditional library services now have the opportunity to get information about all types of subjects, free of political censorship.

As libraries have changed, so, too, has the role of the librarian. Increasingly librarians have assumed the role of educator to teach their users how to find information both in the library and over electronic networks. Public librarians have expanded their roles by providing local community information through publicly accessible computing systems. Some librarians are experts about computers and computer software. Others are concerned with how computer technologies can preserve the human cultural records of the past or assure that library collections on crumbling paper or in old computer files can still be used by people many centuries in the future.

The work of librarians has also moved outside library walls. Librarians have begun to work in the information industry as salespeople, designers of new information systems, researchers, and information analysts. They also are found in such fields as marketing and public relations and in such organizations as law firms, where staffs need rapid access to information.

Although libraries have changed significantly over the course of history, as the following section demonstrates, their cultural role has not. Libraries remain responsible for acquiring or providing access to books, periodicals, and other media that meet the educational, recreational, and informational needs of their users. They continue to keep the business, legal, historical, and religious records of a civilization. They are the place where a toddler can hear his first story and a scholar can carry out her research.

(L.S.E.)

The history of libraries

THE ANCIENT WORLD

In earliest times there was no distinction between a record room (or archive) and a library, and in this sense libraries can be said to have existed for almost as long as records have been kept. A temple in the Babylonian town of Nippur, dating from the first half of the 3rd millennium BC, was found to have a number of rooms filled with clay tablets, suggesting a well-stocked archive or library. Similar collections of Assyrian clay tablets of the 2nd millennium BC were found at Tell el-Amarna in Egypt. Ashurbanipal (reigned 668–*c.* 627 BC), the last of the great kings of Assyria, maintained an archive of some 25,000 tablets, comprising transcripts and texts systematically collected from temples throughout his kingdom.

Ashurbanipal's library

Many collections of records were destroyed in the course of wars or were purposely purged when rulers were replaced or when governments fell. In ancient China, for example, the emperor Shih huang-ti, a member of the Ch'in dynasty and ruler of the first unified Chinese empire, ordered that historical records other than those of the Ch'in be destroyed so that history might be seen to begin with his dynasty. Repression of history was lifted, however, under the Han dynasty, which succeeded the Ch'in in 206 BC; works of antiquity were recovered, the writing of literature as well as record keeping were encouraged, and classification schemes were developed. Some favoured a seven-part classification, which included the Confucian classics, philosophy, rhymed work (both prose and poetry), military prose, scientific and occult writings, summaries, and medicine. A later system categorized writings into four types: the classics, history, philosophy, and miscellaneous works. The steady growth of libraries was facilitated by the entrenchment of the civil service system, founded in the 2nd century during the Han dynasty and lasting into the 20th century; this required applicants to memorize classics and to pass difficult examinations.

Greece and Alexandria. In the West the idea of book collecting, and hence of libraries as the word was understood for several centuries, had its origin in the classical world. Most of the larger Greek temples seem to have possessed libraries, even in quite early times; many certainly had archive repositories. The tragedian Euripides was known as a private collector of books, but the first important institutional libraries in Athens arose during the 4th century BC with the great schools of philosophy. Their texts were written on perishable materials such as papyrus and parchment, and much copying took place. The Stoics, having no property, owned no library; the schools of Plato and of the Epicureans did possess libraries, the influence of which lasted for many centuries. But the most famous collection was that of the Peripatetic school, founded by Aristotle and systematically organized by him with the intention of facilitating scientific research. A full edition of Aristotle's library was prepared from surviving texts by Andronicus of Rhodes and Tyrannion in Rome about 60 BC. The texts had reached Rome as war booty carried off by Sulla when he sacked Athens in 86 BC.

Aristotle's library formed the basis, mainly by means of copies, of the library established at Alexandria, which became the greatest in antiquity. It was planned by Ptolemy I Soter in the 3rd century BC and brought into being by his son Ptolemy II Philadelphus with the collaboration of Demetrius of Phaleron, their adviser. The founders of this library apparently aimed to collect the whole body of Greek literature in the best available copies, arranged in systematic order so as to form the basis of published commentaries. Its collections of papyrus and vellum scrolls are said to have numbered hundreds of thousands. Situated in a temple of the Muses called the Mouseion, it was staffed by many famous Greek writers and scholars, including the grammarian and poet Callimachus (d. *c.* 240 BC), the astronomer and writer Eratosthenes (d. *c.* 194 BC), the philosopher Aristophanes of Byzantium (d. 180 BC), and Aristarchus of Samothrace (d. 145 BC), the foremost critical scholar of antiquity.

Pergamum. In Asia Minor a library rivaling that of Alexandria was set up at Pergamum during the reigns of Attalus I Soter (d. 197 BC) and Eumenes II (d. 160/159 BC). Parchment (*charta pergamena*) was said to have been developed there after the copying of books was impeded by Ptolemy Philadelphus' ban on the export of papyrus from Egypt. (Parchment proved to be more durable than papyrus and so marks a significant development in the history of technical advances in the dissemination of knowledge.) The library was bequeathed with the whole of the kingdom of Pergamum to the Roman people in 133 BC, and Plutarch records an allegation that Mark Antony gave its 200,000 volumes to Cleopatra, to become part of the Alexandrian library.

Rome. There were many private libraries in classical Rome, including that of Cicero. Indeed, it became highly fashionable to own a library, judging from the strictures of the moralizing statesman Seneca and the spiteful jibes by the poet Lucian on the uncultured "book clown." Excavations at both Rome and Herculaneum have revealed what were undoubtedly library rooms in private houses, one at Herculaneum being fitted with bookcases around the walls. A Roman statesman and general, Lucius Licinius Lucullus, who was reckoned one of the richest men in the Roman world at that time and was famous for his luxurious way of life, acquired as part of his war booty an enormous library, which he generously put at the disposal of those who were interested. His biographer, Plutarch, speaks appreciatively of the quality of his book collection, and Cicero tells of visiting the library to borrow a book and finding his friend Cato ensconced there surrounded by books of the Stoic philosophy.

The library planned by Julius Caesar

Julius Caesar planned a public library and entrusted the implementation of his plans to an outstanding scholar and writer, Marcus Terentius Varro, also the author of a treatise on libraries, *De bibliothecis* (which has not survived). Caesar died before his plans were carried out, but a public library was built within five years by the literary patron Asinius Pollio. Describing its foundation in his

Natural History, Pliny coined a striking phrase that has application to libraries generally: *ingenia hominum rem publicam fecit* ("He made men's talents a public possession"). Libraries were also set up by Tiberius, Vespasian, Trajan, and many of the later emperors; the Bibliotheca Ulpia, which was established by Trajan about AD 100 and continued until the 5th century, was also the Public Record Office of Rome.

Byzantium. In the East the library tradition was picked up at Constantinople. It was probably at Caesarea that Constantine I the Great's order for 50 copies of the Christian scriptures was carried out. Under Constantine himself, Julian, and Justinian, the imperial, patriarchal (in the religious sense), and scholarly libraries at Constantinople amassed large collections; their real significance is that for a thousand years they preserved, through generations of uncritical teachers, copyists, and editors, the treasures of the schools and libraries of Athens, Alexandria, and Asia Minor. Losses occurred, but these were mostly due to the habit, noticeable especially in the 9th century, of replacing original texts with epitomes, or summaries. By far the greater part of the Greek classics, however, was faithfully preserved and handed on to the schools and universities of western Europe, and for this a debt is owed to the great libraries and the rich private collections of Constantinople.

Islām and literacy

The Islāmic world. After the death of the Prophet Muḥammad in the 7th century, his followers transcribed his teachings into the Qur'ān, a papyrus codex that quickly became the sacred scripture of the Muslim religion. Believers were encouraged to read it and commit substantial portions to memory. In subsequent decades, as armies of Muḥammad's successors conquered more territory, they took the religion of Islām and a commitment to literacy with them. The establishment of libraries of sacred texts—especially in mosques such as al-Aqṣā in Jerusalem (*c.* 634) and the Great Mosque (Umayyad Mosque) of Damascus (*c.* 721)—was a natural outgrowth of their conquest. Probably drawing inspiration from the Library of Alexandria, the first caliph of the Umayyad dynasty, Muʿāwiyah I, reorganized his personal library in the late 7th century into a prototype that his successors further improved and expanded. Caliph al-Walīd (reigned 705–715) appointed the first so-identified *ṣāḥib al-maṣāḥif* ("curator of books"). By that time the Umayyad collection included hundreds of works on astrology, alchemy, medicine, and military science.

In 750 the ʿAbbāsids seized large portions of the eastern Umayyad empire (Umayyads retained control of the Iberian Peninsula), and under the leadership of al-Manṣūr, the second ʿAbbāsid caliph, many classical Persian and Greek works were translated into Arabic. When Muslims shortly thereafter adopted the technique of papermaking learned from Chinese prisoners of war, they significantly increased their capacity to reproduce the written word cheaply and thus directly affected libraries. By the 10th century Baghdad and Córdoba (still controlled by Umayyads) had developed the largest book markets in the world. Christian monks and scholars were often sent to Córdoba to acquire new works.

Other noteworthy libraries of the Islāmic world include those at Baghdad (under Hārūn ar-Rashīd), Cairo, Alexandria, and also Spain, where there was an elaborate system of public libraries centred on Córdoba, Toledo, and Granada. Arabic works from these libraries began to reach Western scholars in the 12th century, about the time that Greek works from Constantinople were filtering through to the West.

THE MIDDLE AGES AND THE RENAISSANCE

The role of the European monasteries. As European monastic communities were set up (from as early as the 2nd century AD), books were found to be essential to the spiritual life. The rule laid down for observance by several monastic orders enjoined the use of books: that of the Benedictine order, especially, recognized the importance of reading and study, making mention of a "library" and its use under the supervision of a precentor, one of whose duties was to issue the books and take daily inventory of them. Scriptoria, the places where manuscripts were copied out, were a common feature of the monasteries—again, especially in those of the Benedictine order, where there was a strict obligation to preserve manuscripts by copying them. Many—Monte Cassino (529) and Bobbio (614) in Italy; Luxeuil (*c.* 550) in France; Reichenau (724), Fulda (744), and Corvey (822) in Germany; Canterbury (597), Wearmouth (674), and Jarrow (681) in England—became famous for the production of copies. Rules were laid down for the use of books, and curses invoked against any person who made off with them. Books were, however, lent to other monasteries and even to the secular public against security. In this sense, the monasteries to some extent performed the function of public libraries.

"The Scriptorium in the Bell Tower of San Salvador at Tavara," from the Beatus of Liebana's Commentary on the Apocalypse (M. 429 folio 183), 1220. In the Pierpont Morgan Library, New York.

By courtesy of the Pierpont Morgan Library, New York

Contents of monastic libraries

The contents of these monastic libraries consisted chiefly of the scriptures, the writings of the early Church Fathers and commentaries on them, chronicles, histories such as Bede's *Historia ecclesiastica gentis Anglorum* ("Ecclesiastical History of the English People"), philosophical writings such as those of Anselm, Peter Abelard, St. Thomas Aquinas, and Roger Bacon, and possibly some secular literature represented by the Roman poets Virgil and Horace and the orator Cicero. After the universities were founded, beginning in the 11th century, monkish students, on returning to their monasteries, deposited in the libraries there the lecture notes they had made on Aristotle and Plato, on law and medicine, and so forth, and in this way expanded the libraries' contents.

The new learning. In Europe the libraries of the newly founded universities—along with those of the monasteries—were the main centres for the study of books until the late Middle Ages; books were expensive and beyond the means of all but a few wealthy people. The 13th, 14th, and 15th centuries, however, saw the development of private book collections. Philip the Good, duke of Burgundy, and the French kings Louis IX and Charles V (who may be looked upon as the founder of the Bibliothèque du Roi ["King's Library"], which later became the Bibliothèque Nationale ["National Library"] in Paris) were great collectors, as were also such princes of the church as Richard de Bury, bishop of Durham (d. 1345), who wrote a famous book in praise of books, *Philobiblon* (*The Love of Books;*

first printed in Cologne, 1473). But new cultural factors—including the growth of commerce, the new learning of the Renaissance (which was based on newly discovered classical texts), Johannes Gutenberg's invention of a printing press using movable type, and a substantial expansion of lay literacy—widened the circle of book collectors to include wealthy merchants whose libraries contained herbals, books of law and medicine, and books of hours and other devotional works. Italian humanists, such as Petrarch and Giovanni Boccaccio, searched for and copied manuscripts of classical writings (such as those of Cicero and Tacitus) to establish their scholarly libraries. The scholars Niccolò Niccoli (librarian to Cosimo de' Medici, the 14th-century ruler of Florence and a considerable patron of the arts) and Gian Francesco Poggio Bracciolini shared this enthusiasm for the classics and ransacked Europe and the Middle East for manuscripts of the writers of Greece and Rome. Notable collections of books were made outside Italy, too (though Florence remained the centre of the rising book trade): by Diane de Poitiers, mistress of Henry II of France; by Jean Grolier, a high French official and diplomat, who was a great patron of bookbinders; by John Tiptoft, earl of Worcester; by Henry VII and Henry VIII of England; and by many others.

Renaissance public libraries

On the basis of Niccoli's library, Cosimo de' Medici set up the Biblioteca Marciana in Florence in the convent of San Marco. The rich library of Lorenzo the Magnificent, grandson of Cosimo and an even greater patron of learning and the arts, also became a public library. It was opened in 1571 in a fine building designed by Michelangelo and still exists as the Biblioteca Laurenziana (though in 1808 it was amalgamated with the Marciana to form the Biblioteca Medicea-Laurenziana [Medicean-Laurentian Library]). Many other princely libraries were formed at this time, including that of Matthias I (Matthias Corvinus) of Hungary and the library of the Escorial in Madrid (founded 1557), based on the collections of Philip II. The Vatican library also dates its foundation from this time.

Effects of the Reformation and religious wars. In England the end of the monastic libraries came in 1536–40, when the religious houses were suppressed by Henry VIII and their treasures dispersed. No organized steps were taken to preserve their libraries. Even more wholesale destruction came in 1550: Henry VIII and Edward VI aligned with the "new learning" of the humanists; and university, church, and school libraries were purged of books embodying the "old learning" of the Middle Ages. The losses were incalculable. During Elizabeth's reign, however, the archbishop of Canterbury, Matthew Parker, and Elizabeth's principal adviser, William Cecil,

By courtesy of the Dean and Chapter of Hereford Cathedral, England; photograph, F.C. Morgan

The chained library of the Cathedral Church of the Blessed Virgin Mary and St. Aethelberht, Hereford, Eng.

took the lead in seeking out and acquiring the scattered manuscripts. Many other collectors were also active, including Sir Robert Cotton and Sir Thomas Bodley. As a result, a considerable portion of the libraries that had been scattered at the suppression was, by 1660, reassembled in collections—Parker's eventually went to Corpus Christi College at Cambridge; Cotton's to the British Museum library, which now forms part of the British Library; and Bodley's to form the Bodleian Library at Oxford.

Elsewhere in Europe, the period of the Reformation also saw many of the contents of monastic libraries destroyed, especially in Germany and the northern countries. The Reformation leader Martin Luther, however, did himself passionately believe in the value of libraries, and in a letter of 1524 to all German towns he insisted that neither pains nor money should be spared in setting up libraries. As a consequence, many town libraries in Germany, including those at Hamburg (1529) and Augsburg (1537), date from this time. These, and the libraries of the newly created universities (such as those of Königsberg [now Kaliningrad, Russia], Jena, and Marburg), were partly, at any rate, built up on the basis of the old monastic collections. In Denmark, similarly, some books from the churches and monasteries were incorporated with the new university library, though many were destroyed.

Libraries in Germany suffered severely in the Thirty Years' War. The Bibliotheca Palatina at the University of Heidelberg (founded 1386), for example, was taken as the spoil of war by Maximilian I of Bavaria, who offered it to Pope Gregory XV in 1623; and Gustavus Adolphus sent whole libraries to Sweden, most of them to swell the library of the University of Uppsala, which he had founded in 1620. The collections of the Royal Library in Stockholm were similarly enriched by the war booty that fell to Sweden during the reigns of Queen Christina and Charles X. In France, Italy, southern Germany, and Austria, where the Roman Catholic faith remained unshaken, the old libraries remained and were supplemented by new ones set up for educational purposes by the Society of Jesus (the Jesuits).

The Islāmic world. Like the European monastic libraries, book collections in the Islāmic countries at first were attached to religious institutions, both mosques and *madrasahs* (the theological and law schools centred on study of the Qur'ān). Scholars donated their personal collections to mosques, which usually kept only the religious books, sometimes setting up an adjunct library in which the books of a more secular nature were placed. These secular collections were open to the public. Apart from the libraries associated with mosques, there were many large collections housed in palaces and the homes of the wealthy. Notable libraries were established by the ʿAbbāsid caliph al-Maʾmūn in Baghdad in the 9th century and by the Fāṭimid caliph al-Mustanṣir in 11th-century Cairo. Typical private and public collections usually included regional histories and works of geography, travel, astrology, and alchemy.

17TH AND 18TH CENTURIES AND THE GREAT NATIONAL LIBRARIES

In the 17th and 18th centuries book collecting everywhere became more widespread. The motive sometimes was sheer ostentation, but often it was genuine love of scholarship. Throughout Europe and in North America, several fine private collections were assembled, many of which were eventually to become the core of today's great national and state libraries—for this was also the period that saw the establishment of new national and university collections.

Parish, town, and subscription libraries

There were, of course, other developments. In England there were established a number of parish libraries, attached to churches and chiefly intended for the use of the clergy (one of the earliest, at Grantham in Lincolnshire, was set up as early as 1598, and some of its original chained books are still to be seen there). They were sometimes the result of lay donation: a Manchester merchant, Humphrey Chetham, left money in 1653 for the foundation of parish libraries in Bolton and Manchester and also for the establishment of a town library in Manchester (which still exists, housed in its original bookcases, in its original building). Later, in the 18th century, especially in England (though also elsewhere in Europe) and the United States, there was a great vogue for the circulating and subscription libraries—societies that provided reference service and lending collections for their members and had much influence on the formation of popular literary taste, especially in fiction.

Library planning. The private libraries of powerful and influential collectors, such as Cardinal Mazarin in France, were so large that a new approach to library organization was needed. The Escorial library in Madrid, erected in 1584, had been the first to do away with the medieval book bays, which were set at right angles to the light source, and to arrange its collection in cases lining the walls. The old practice of chaining books to their cases was gradually abandoned; and the change to the present arrangement, standing books with their spines facing outward, began in France—probably with the personal library of the lawyer, councillor of state, historian, and bibliophile Jacques-Auguste de Thou (d. 1617). Mazarin's library was in the charge of Gabriel Naudé, who produced the first modern treatise on library economy, *Advis pour dresser une bibliothèque* (1627; *Advice on Establishing a Library*). This work marked the transition to the age of modern library practice. One of its first fruits was the library of the diarist Samuel Pepys; in the last 14 years of his life Pepys devoted much time to the organization of his collection, and he left it to Magdalene College, Cambridge.

Naudé's concept of a scholarly library, systematically arranged, displaying the whole of recorded knowledge and open to all scholars, took root. It was above all absorbed by the philosopher Gottfried Wilhelm Leibniz (1646–1716), a prominent librarian of his age, who conceived the idea of a national bibliographical organization that would provide the scholar with easy access to all that had been written on his subject.

Emergence of national collections. The scope of European scholarship and inquiry expanded rapidly during the 17th and 18th centuries, especially in the field of historical studies and in philosophy. In France, de Thou, highly qualified as a collector, was made director in 1593 of the Bibliothèque du Roi (founded by Charles V and largely reorganized during the 15th century by Louis XII). Mazarin's library was scattered when he was compelled to leave France during the period of unrest known as the Fronde, but it was reassembled when he returned to power in 1661. Rehoused in a new building, it was opened to the public in 1691. It remained one of France's great libraries until after the French Revolution, when it was incorporated with other collections (including the Bibliothèque du Roi) to form the Bibliothèque Nationale, today one of the world's great libraries. August, Duke von Braunschweig, established a library in 1604 that later became the Herzog August Bibliothek at Wolfenbüttel, one of the finest libraries in Europe (Leibniz was its librarian from 1690 to 1716). A library assembled by the elector Friedrich Wilhelm of Brandenburg was founded in 1659 and later became the Prussian State Library. The collections of the English book collectors Sir Hans Sloane, Sir Robert Cotton, and Edward and Robert Harley, earls of Oxford, formed the basis of the British Museum collection (1753), which was enlarged in 1757 by the addition of the Royal Library, containing books collected by the kings of England from Edward IV to George II.

The great national collections

The effects of the French Revolution. On the continent of Europe the anticlerical movement that found expression in revolution sealed the fate of many monastic and church libraries: those in France, for example, were expropriated in 1789; in Germany in 1803; in Spain in 1835. In France books were collected in the main towns of the *départements* in what were called *dépots littéraires.* In 1792 the same fate befell the collections of aristocratic families, and these, too, were added to the *dépots.* The enormous accumulations caused problems, and many books were lost, but the plan of coordinating library resources throughout the country was carried out. The Bibliothèque Nationale received some 300,000 volumes, and new libraries were set up in many important provincial cities. In Bavaria the state library was greatly enriched by the contents of more than 150 confiscated libraries, and many of the provincial libraries were similarly enlarged. In Austria, as a result of confiscations, *Studienbibliotheken* (study libraries) were set up at Linz, Klagenfurt, and Salzburg, the university libraries at Graz and Innsbruck were substantially enlarged, and many valuable acquisitions accrued to the Hofbibliothek in Vienna.

The library of Samuel Pepys at Magdalene College, Cambridge.
By courtesy of the Master and Fellows of Magdalene College, Cambridge University

LATER DEVELOPMENTS

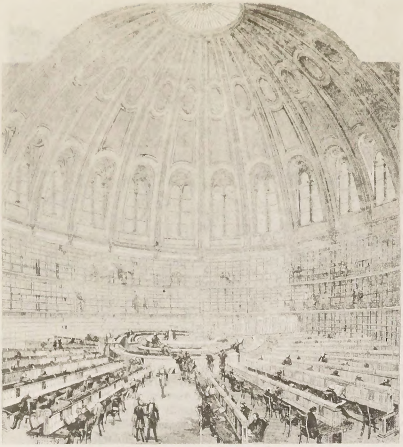
The difficulties of library management grew in the 19th century. Libraries had increased in size, but their growth had been haphazard; administration had become weak, standards of service almost nonexistent; funds for acquisition tended to be inadequate; the post of librarian was often looked on as a part-time position; and cataloging was frequently in arrears and lacked proper method.

The university library at Göttingen was a notable exception. Johann Gesner, the first librarian, working in close association with the curator of the university, G.A. von Münchhausen, and proceeding on the principles laid down by Leibniz, made strenuous efforts to cover all departments of learning; the library provided good catalogs of carefully selected literature and was available to all as liberally as possible. The library's next director, C.G. Heyne, enthusiastically followed the same principles, with the result that Göttingen became the best-organized library in the world.

A leading figure in the transformation of library service was Antonio (later Sir Anthony) Panizzi, a political refugee from Italy who began working for the British Museum in 1831 and was its principal librarian from 1856 to 1866. From the start he revolutionized library administration, demonstrating that the books in a library should match its declared objectives and showing what these objectives should be in the case of a great national library. He perceived the importance of a good catalog and to this end elaborated a complete code of rules for catalogers. He also saw the potential of libraries in a modern community as instruments of study and research, available to all, and, by his planning of the British Museum reading room and its accompanying bookstacks, showed how this potential might be realized. His ideas long dominated library thought in the field of scholarly—or, as they are now called, research—libraries and achieved major expression in the Library of Congress in Washington, D.C.

The revolution in library administration

By the middle of the 19th century the idea had been accepted that community libraries might be provided by local authorities at public expense. This proved a significant stage in the development of library provision. Panizzi had stated that he wanted the facilities of a great library to be available to poor students so that they could indulge their "learned curiosity"; in England in 1850 an act of



British Museum reading room designed by Sir Anthony Panizzi, 1854. Illustration by Sydney Smirke, from *The Illustrated London News*, 1857.

By courtesy of the trustees of the British Museum; photograph, J.R. Freeman & Co. Ltd.

Parliament was passed enabling local councils to levy a rate for the provision of free library facilities.

The paradigm for libraries and librarianship shifted radically in the 20th century with the advent of new information technologies. By the end of the century, computer-based systems had given individuals access to an enormous network of information. Especially in the world's major urban centres, the library's traditional means of sharing access to information, such as the owning and lending of books and other materials or the sharing of these resources with sister libraries, were increasingly supplanted by the use of electronic databases that contained everything from library catalogs and subject area indexes and abstracts to journal articles and entire book-length texts. As individuals using home computers became familiar with a worldwide electronic network, the library as a storehouse site was challenged by the so-called virtual library, accessible by computer from any place that had telephone or cable lines. The role of the professional librarian also evolved, as many were called upon to be familiar with and to train others to use a variety of electronic databases. (Ed.)

Types of libraries

Library services available throughout the world vary so much in detail from country to country that it is difficult to present anything but the most general picture of their activities. Nevertheless, they follow a broad but discernible pattern that has evolved over the years.

NATIONAL LIBRARIES

Aims of a national library

In most countries there is a national or state library or a group of libraries maintained by national resources, usually bearing responsibility for publishing a national bibliography and for maintaining a national bibliographical information centre. National libraries strive principally to collect and to preserve the nation's literature, though they try to be as international in the range of their collections as possible.

Most national libraries receive, by legal right (known in English as legal, or copyright, deposit), one free copy of each book and periodical printed in the country. Certain other libraries throughout the world share this privilege, though many of them receive their legal deposit only by requesting it.

The Bibliothèque Nationale in Paris, the British Library in London, and the Library of Congress in Washington, D.C., are among the most famous and most important national libraries in the Western world. Their importance springs from the quality, size, and range of their collections, which are comprehensive in scope, and from their attempts to maintain their comprehensiveness. They achieve the latter quality with diminishing success in view of the vastly increased number of publications that daily appear throughout the world, the failure of publishers to provide legal-deposit copies, and the difficulty of ensuring adequate representation of publications issued in the developing countries.

Bibliothèque Nationale. As indicated above, the Bibliothèque Nationale before the French Revolution was known as the Bibliothèque du Roi and owes its origin to Charles V. During the 15th and 16th centuries it received a number of important collections of manuscripts; in 1617, under the librarianship of de Thou, its right to legal deposit was reaffirmed and continued to be rigidly enforced. In the first quarter of the 18th century, four of the library's departments (of prints, coins, printed books, and manuscripts) were created; it was opened to the public in 1735. Enormous additions accrued to the library as a result of the Revolution and the confiscation of aristocratic and church private collections. The catalog of the library on cards was completed under the librarianship (1874–1905) of Léopold Delisle, and in 1897 he made a start to the task of compiling a printed catalog in volume form.

The present-day Bibliothèque Nationale plays a leading role in the French national library service. Its Directorate of Libraries oversees all public libraries and participates in the training of library professionals.

The British Library. For more than two centuries the British Museum combined a great museum of antiquities with a great comprehensive library. The library was founded in 1753 by the acceptance of the bequest of the collections of Sir Hans Sloane, physician to King George II and president of the Royal Society. The library was built up on the basis of two other important collections, that of Sir Robert Cotton and that of Edward and Robert Harley, earls of Oxford; to these were added the Royal Library, given by George II in 1757. With this collection came also the right to legal deposit of one copy of every book published in the British Isles; this right is generally enforced, yet many titles arrive only slowly and some not at all. These four basic collections were notably enlarged during the first century of the library's history by the addition of many private collections, including the libraries of King George III (1823) and of Thomas Grenville (1846). The library's printed catalog, executed under the guidance of Sir Anthony Panizzi, was issued between 1881 and 1905.

British Library Act of 1972

The British Museum's library was separated from the museum under the British Library Act of 1972 and by July 1, 1973, was reorganized as the British Library Reference Division. The British Library Lending Division was formed from the amalgamation of two previously existing libraries: the National Central Library, which had been the centre for interlibrary lending since 1927 and which had a collection of some 400,000 books and periodicals, mainly in the humanities and social sciences; and the National Lending Library for Science and Technology, which had been opened in 1962 by the Department of Scientific and Industrial Research.

The British Library Bibliographic Services Division was formed from the British National Bibliography Ltd., an independent organization set up in 1949 to publish a weekly catalog of books published in the United Kingdom and received at the British Museum by legal deposit. The *British National Bibliography*, as this weekly catalog was called, quickly established itself as a foremost reference work, both for book selection and cataloging and for reference retrieval. After the reorganization of 1973 the division expanded the computerizing of current cataloging and the central provision of both printed cards and machine-readable entries. The BLAISE service (British Library Au-

The new Bibliothèque Nationale on the Left Bank of the Seine, Paris. The four towers were designed to look like open books standing on end.

© Annebicque/SYGMA

tomated Information Service) offers a cataloging facility to any library wishing to participate, and the Bibliographic Services Division and its predecessor, the British National Bibliography, cooperated closely with the U.S. Library of Congress in the Project for Machine-Readable Cataloging (MARC), which provides on-line access to the catalogs of the current acquisitions of the British Library Reference Division and the Library of Congress.

Library of Congress. The U.S. Library of Congress in Washington, D.C., is probably the largest national library, and its collection of modern books is particularly extensive. It was founded in 1800 but lost many books by fire during a bombardment of the Capitol by British troops in 1814. These losses were to some extent made good by the purchase of Thomas Jefferson's library shortly thereafter. The library remained a strictly congressional library for many years, but, as the collections were notably enlarged by purchases and by additions under the copyright acts, the library became and remained—in effect, although not in law—the national library of the United States. The public has access to many of the collections.

Through a service begun by Herbert Putnam, librarian from 1889 to 1939, the Library of Congress makes its catalog available to many thousands of subscribing American libraries and institutions.

The library's impact on librarianship has always been of the highest value. Through the Library of Congress Classification, the printed catalog cards, and MARC (see below *Technical services*), the library's practices are widely followed. Its last great printed product was the 754-volume *National Union Catalog: Pre-1956 Imprints.* In 1983 the library began producing most of the *National Union Catalog* on microfiche (sheets of microfilm containing rows of microimages of pages of printed matter). It serves as a centralized bureau for information on the acquisition of materials worldwide and distributes cataloging data to other libraries. It also has taken a considerable role in the areas of materials preservation and the research and development of new methods of information storage.

Russian State Library. Of a size and importance comparable to the Library of Congress, the Russian State Library (formerly called the Lenin Library) in Moscow is the national library of Russia. It receives several copies of all publications from throughout the country and distributes copies to specialist libraries. It issues printed cards for the *Bibliography of Periodicals, 1917–1947* and for a cooperative catalog that lists the holdings of the Russian State Library, the Saltykov-Shchedrin Public Library in St. Petersburg, the Library of the Russian Academy of Sciences also in St. Petersburg, and the Central Book Office. It organizes domestic and international lending and exchanges and offers courses of lectures for professional education and also for readers. It formerly produced the Soviet Library-Bibliographical Classification scheme based on a Marxist-Leninist classification of knowledge.

Other national collections. There are many other national libraries with important collections and very long histories. The Bibliothèque Royale Albert I in Brussels, founded in 1837 and centred on the 15th-century collection of the dukes of Burgundy, is the national library of Belgium and the centre of the country's library network; it maintains a regular lending service with the university libraries and with the large town library of Antwerp. The Dutch Royal Library in The Hague was founded in 1798, and it, too, is the centre of a well-developed interlibrary loan system. Because the unification of Italy in the 19th century brought together many city-states that had major libraries, the country has a number of national libraries, the chief being the Biblioteca Nazionale Centrale Vittorio Emanuele II in Rome, founded in 1875, and the historically richer Biblioteca Nazionale Centrale at Florence, founded in 1747. Other Italian national libraries are at Milan, Naples, Palermo, Turin, and Venice. Germany was equally remarkable before World War II both for the importance of its state or provincial libraries and for the lack of a recognized national library. The former Preussische Staatsbibliothek was given national status in 1919. That library became East Germany's national library after World War II. In 1990, after the reunification of Germany, the Deutsche Bibliothek in Frankfurt am Main was merged with the Deutsche Bücherei in Leipzig and the Deutsche Musikarchiv to form the national library of Germany. The Austrian National Library, founded by the emperor Maximilian I in 1493, has rich collections—notably of manuscripts from the Austrian monasteries and from the library of Matthias I Corvinus, dispersed after the capture of his capital, Buda, by the Turks in 1526. The National Library of Australia in Canberra, formally created by legislation in 1960, grew out of the Commonwealth Parliamentary Library, established in 1901.

National libraries in Asia

The National Library of China in Peking consists of the books and archives from imperial libraries dating to the Southern Sung dynasty (founded 1127). It also contains

© Christopher Liu/ChinaStock Photo Library

A reading area in the National Library of China in Peking.

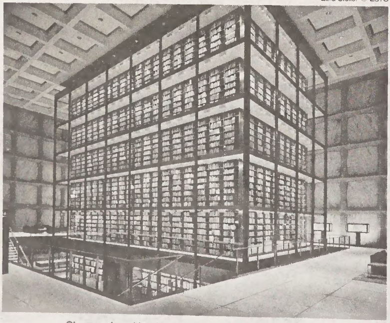
An atrium and (left) reading rooms of the National Diet Library in Tokyo. The tapestry at right is the work of Ikeda Masuo.
© David Ladd Nelson

inscribed tortoise shells and bones, ancient manuscripts, and block-printed volumes, as well as books from the Ch'ing dynasty, imperial colleges, and private collectors. The National Diet Library (1948) in Tokyo counts among its holdings some four million volumes. Based on the collections of the former Imperial Library (1872), it is organized like the U.S. Library of Congress and publishes a computer-generated national bibliography. The National Library of India (formerly the Imperial Library) in Calcutta was founded in 1903. It is the largest library in India and holds a fine collection of rare books and manuscripts. In some countries, such as Iceland, Norway, and Israel, the national library is combined with a university library.

UNIVERSITY AND RESEARCH LIBRARIES

Before the invention of printing, it was common for students to travel long distances to hear famous teachers. Printing made it possible for copies of a teacher's lectures to be widely disseminated, and from that point universities began to create great libraries. The Bodleian Library (originally established in the 14th century) at Oxford University and Harvard University Library (1638) at Cambridge, Mass., are superior to many national libraries in size and quality. In addition to a large central library, often spoken of as the heart of a university, there are often smaller, specialized collections in separate colleges and institutes. The academies of science in Russia and various other former Soviet republics and those in the countries of eastern Europe consist of groups of specialized institutes, and, while not all act as universities in awarding degrees, their research function has the same significance. Some, as in Hungary and Romania, serve as the national library.

Ezra Stoller © ESTO



Glass-enclosed bookstack of the Beinecke Rare Book and Manuscript Library at Yale University, New Haven, Conn.

In a university library many users may seek to use the same books at the same time. The difficulty of providing multiple copies has vexed most university librarians, who must balance slender resources against sometimes vociferous demand. To handle the problem, many libraries have set up a short-loan collection (typically called the reserve collection) from which books may be borrowed for as little as a few hours. The use of computers for circulation control has brought some relief through great flexibility of operation and capacity for instant recall of information on the whereabouts of a particular work.

Cooperation among libraries

The range of research carried out at a traditional university may encompass every aspect of every discipline, and even the largest university libraries have long recognized the need for cooperation with others, first in cataloging and later in acquisitions. Automation and computers have helped, too, by making it possible for readers in one library to consult the catalogs of others, as well as independent databases, indexes, and abstracts, by means of computer networks. The printing of multiple volumes of union catalogs, especially for periodicals, proved the value to scholars of sharing information on catalogs and collections. Many universities have made available catalogs of their special collections and have arranged for the reproduction both of rare individual works and of complete collections on microfilm and in other formats. An example is the Goldsmiths'-Kress collection of early works in economics, which combines the holdings of the Goldsmiths' Library at the University of London and the Kress Library at Harvard.

PUBLIC LIBRARIES

Role of public libraries

Public libraries are now acknowledged to be an indispensable part of community life as promoters of literacy, providers of a wide range of reading for all ages, and centres for community information services. Yet, although the practice of opening libraries to the public has been known from ancient times, it was not without considerable opposition that the idea became accepted, in the 19th century, that a library's provision was a legitimate charge on public funds. It required legislation to enable local authorities to devote funds to this cause.

Public libraries now provide well-stocked reference libraries and wide-ranging loan services based on systems of branch libraries. They are further supplemented by traveling libraries, which serve outlying districts. Special facilities may be provided for the old, the blind, the hearing-impaired, and others, and in many cases library services are organized for local schools, hospitals, and jails. In the case of very large municipalities, library provision may be on a grand scale, including a reference library, which has many of the features associated with large research libraries. The New York Public Library, for example, has rich collections in many research fields; and the Boston Public Library, the first of the great city public libraries in the United States (and the first to be supported by direct public taxation), has had from the first a twofold character as a library for scholarly research as well as for general reading. In the United Kingdom the first tax-supported public libraries were set up in 1850; they provide a highly significant part of the country's total national library service. The importance of public library activities has been recognized in many countries by legislation designed to ensure that good library services are available to all without charge.

In many cases public libraries build up collections that relate to local interests, often providing information for local industry and commerce. It is becoming more usual for public libraries to lend music scores, phonograph records, compact discs, and, in some countries—notably Sweden and the United Kingdom—original works of art for enjoyment, against a deposit, in the home.

Not all countries provide public library services of an equally high standard, but there has been a tendency to recognize their value and to improve services where they exist or to introduce them where they do not. Public librarians work strenuously, through such organizations as the International Federation of Library Associations (IFLA), for such developments.

SPECIAL LIBRARIES

The national, university, and public libraries form the network of general libraries more or less accessible to the general public. They take pride in special collections, which are built around a special subject interest. Beyond this network are a large number of libraries established by special groups of users to meet their own needs. Many of these originated with learned societies and especially with the great scientific and engineering societies founded during the 19th century to provide specialist material for their members. Thus some special libraries were founded independently of public libraries and before major scientific departments were developed in national libraries; for example, the National Reference Library of Science and Invention, now the Science Reference Library and part of the British Library, was originally established at the U.K. Patent Office.

With the coming of the Industrial Revolution arose the need for a working class educated in technology, and industrialists and philanthropists provided facilities and books of elementary technical instruction. In the United Kingdom the Mechanics' Institutes were founded in the rapidly growing industrial towns to provide books and lectures to workers and tradesmen at prices lower than those of the subscription libraries.

Functions and services of specialist libraries

Special libraries are frequently attached to official institutions such as government departments, hospitals, museums, and the like. For the most part, however, they come into being in order to meet specific needs in commercial and industrial organizations. Special libraries are planned on strictly practical lines, with activities and collections carefully controlled in size and scope, even though these libraries may be and in fact often are large and wide-ranging in their activities; they cooperate widely with other libraries. They are largely concerned with communicating information to specialist users in response to—or preferably in anticipation of—their specific needs. Special libraries have therefore been much concerned with the theoretical investigation of information techniques, including the use of computers for indexing and retrieval. It was in this area that the concept of a science of information flow and transfer emerged as a new field of fundamental theoretical study. The concept underpins the practices not only of special libraries but of all types of library and information services.

SCHOOL LIBRARIES

Where public libraries and schools are provided by the same education authority, the public library service may include a school department, which takes care of all routine procedures, including purchasing, processing with labels, and attaching book cards and protective covers; the books are sent to the schools ready for use. This is done in Denmark and in some parts of the United Kingdom. In other countries—the United States, for example—processing may be contracted out to a specialist supplier. In most countries, in fact, school and public libraries cooperate closely.

Teachers who take an interest in the school library make a considerable contribution to its progress, and many have acquired qualifications in librarianship, recognizing that a modern library requires full-time attention and a variety of skills. The school librarian must have a close knowledge of and sympathy with the work of the teaching staff. School libraries have been the scene of significant research and experiment with many different media, so much so that some school libraries have become resource centres. Teachers accustomed to using visual aids, often indeed to making their own, have come to expect the library to provide such materials as collections of photographs, slides, films and filmstrips, videotapes, and artifacts for work in subjects such as history and mathematics. Some school librarians use the term "realia" to describe these resources.

PRIVATE LIBRARIES

The libraries owned by private individuals are as varied in their range of interest as the individuals who collected them, and so they do not lend themselves to generalized treatment. The phrase *private library* is anyway unfortunate because it gives little idea of the public importance such libraries may have. Private collectors are often able to collect in depth on a subject to a degree usually impossible for a public institution; being known to booksellers and other collectors, they are likely to be given early information about books of interest to them; they can also give close attention to the condition of the books they buy. In these ways they add greatly to the sum of bibliographical knowledge (especially if they make their collections available to scholars).

Henry Clay Folger, for example, collected no fewer than 70 copies of one book—the first collected edition of Shakespeare's plays. (In 1932 he opened the Folger Shakespeare Library in Washington, D.C., which had been built to house his collection.) As a result of his collecting he added greatly to the sum of knowledge about the printing of Shakespeare's plays and about 17th-century printing in general. Collectors of private libraries have sometimes benefited posterity by leaving their collections to public institutions or founding a library. Examples in the United States include Henry E. Huntington, John Carter Brown, William L. Clements, and J.P. Morgan. The tradition has long been established in Europe, where many important libraries have been built up around the nucleus of a private collection.

© Peter Aaron/ESTO

The main reading room in the Folger Shakespeare Library, Washington, D.C., a private library established by Henry Clay Folger in 1932.

SUBSCRIPTION LIBRARIES

Famous subscription libraries

Part public, part private, these libraries enjoyed much popularity from the late 17th to the 19th century. Many of them were set up by associations of scholarly professional groups for the benefit of academies, colleges, and institutions, but their membership was also open to the general public. Some of them are still in existence: perhaps the most famous are the Library Company of Philadelphia, founded by Benjamin Franklin in 1731; the Boston Athenaeum, founded in 1807; and the London Library, opened largely at the request of Thomas Carlyle in 1841, which today has a wide-ranging collection for loan to its members in their homes.

During the 19th century, the great size of many subscription libraries enabled them to wield much influence over publishers and authors: Mudie's Circulating Library, for instance, established in London in 1842, would account for the sale of as much as 75 percent of a popular novel's edition. Nevertheless, these libraries were for the most part unable to survive, and the service they gave is now largely provided by the free public libraries.

ARCHIVES

Archives are collections of papers, documents, and photographs (often unpublished or one-of-a-kind), and sometimes other materials that are preserved for historical reasons. They are created in the course of conducting business activities of a public or private body. Until the mid-15th century and the use of the printing press, such records were not distinguished from library materials and were preserved in the same places as other manuscripts. The importance now accorded to public records has been recognized as one outcome of the French Revolution, when for the first time an independent national system of archive administration was set up, for whose preservation and maintenance the state was responsible and to which there was public access.

While the administration of archives shares with libraries the basic obligation to collect, to preserve, and to make available, it has to employ different principles and management techniques. Libraries might be described as collecting agencies, whereas archival institutions are receiving agencies: they do not select—their function is to preserve documents as organic bodies of documentation. They must respect the integrity of these bodies of documents and maintain as far as possible the order in which they were created. And, of course, the documents need catalogs and finding aids, or guides.

Distinction between public and private archives

A distinction has to be drawn between public and private archives. Every state, broadly speaking, now recognizes the need to preserve its own official records and is expected to maintain a system of archive administration, which has the function of collecting them, preserving them, and making them publicly available after the appropriate lapse of time. Among the best known are the Archives Nationales in France, the U.S. National Archives, and the British Public Record Office. Nonofficial archives—the records of the day-to-day activities of an institution or a business—are now recognized as having great value for socioeconomic history, and they are frequently sought by libraries for their historical value and preserved in manuscript and similar collections. It is the practice of many institutions, such as universities, professional and commercial organizations, and ecclesiastical establishments, to set up their own archive departments. (F.C.F./D.J.F./Ed.)

The library operation

TRAINING AND LIBRARY MANAGEMENT

Throughout the centuries, librarians have preserved books and records from the hazards of war, fire, and flood, and it is no idle boast to say that they have played a large part in maintaining the cultural heritage of their countries. Although the traditional librarian acted primarily as a keeper of records, the concept of an active service of advice and information eventually appeared as a legitimate extension of the role of custodian.

The rise of scientific and industrial research and the establishment of public libraries in the 19th century led to the greatly increased emphasis on the subject approach and the role of systematic cataloging and classification in addition to the accepted function of building the collection and the consequent need for expert knowledge of bibliography, both systematic and analytic. In the industrial library in particular, the information officer was almost entirely concerned with the information contained on documents and was indifferent to their form; in this scheme a scrap of paper recording an important telephone call would have more significance than an incunabulum (a book printed before 1501). The proliferation of different forms of record eventually led to a much wider view of information storage and retrieval methods, often requiring the intervention of subject specialists who understood the work of their specialist colleagues.

The professional librarian. Now sometimes known as information specialists, librarians often specialize in certain areas. Their professional skills range from those of the archivist, who is concerned with records management, records appraisal, accessioning and arrangement, archival buildings and storage facilities, preservation and rehabilitation, and reference services (including exhibition and publication), to those of the information scientist, who is concerned with research on the nature of information itself and the process of information flow and transfer between individuals and communities. The various branches of the information profession share many objectives, practices, and skills. Each branch works to make the records of human progress readily available, and the contribution of each to society can only suffer from the lack of integration into a larger whole.

The personnel requirements of the profession include several categories, based on various kinds of specialist knowledge and skills. These include a knowledge of the nature of documents and their role in collection building, skills in the organization of knowledge through cataloging and classification, an ability to analyze and survey needs and to disseminate information in response to and in advance of inquiries, and, often, a high level of computer literacy. Support personnel are needed to maintain the equipment, both hardware and software, and clerks, technicians, and stewards also are essential.

Training institutes. Most of the initiatives for the education and training of professionals have come from librarians or their professional associations. In the United States the first university school for librarians was established in 1887 by Melvil Dewey at Columbia University. The American Library Association (ALA) pursued a policy of accreditation in an effort to ensure that library schools offering a professional qualification meet the standards established by the profession itself. The first British library school was established in University College, London, in 1919, and until 1946 all other qualifications were gained through public examinations that were conducted by the Library Association. Today there are many other schools, most in polytechnic institutes, where the Library Association's own standards continue to influence the curriculum. The association's successive syllabi have had considerable importance for countries such as Ghana, Nigeria, and the Caribbean states.

In continental Europe most professional education takes place in universities and similar institutions of higher learning. The University of Budapest (now Loránd Eötvös University) in Hungary began courses in the Faculty of Philosophy in 1949, and in 1964 a senior-level course in documentation was organized jointly by the university's Chair of Library Science and the National Technical Library and Documentation Centre. In the Czech Republic, library and information science courses are given at the Chair of Library Science and Scientific Information in Charles University. Slovakia's library courses are taught by the Faculty of Philosophy and Letters of Comenius University in Bratislava. In France the long-established École Nationale des Chartes, which mainly trains archivists, also prepares students for the public, national, and university libraries. The École Nationale Supérieure des Bibliothèques belongs to the Direction des Bibliothèques, and the École de Bibliothécaires-Documentalistes is a private institution of the Institut Catholique de Paris.

China's Peking and Wu-han universities have advanced courses and research programs in librarianship, and professional qualifications may also be gained by correspondence. In 1985, with the help of the United Nations Educational, Scientific and Cultural Organization (UNESCO) and the British Council, a master's degree course in information studies was begun at the Institute for Scientific and Technical Information in China.

Training once weighted heavily toward historical and bibliographic aspects of library management has since been balanced with more emphasis on scientific literature, indexing and abstracting techniques, and information technology. Much more research effort is now directed also to the theory of information transfer and the development of mathematical models for this and to other aspects of management in library and information services.

LIBRARY MATERIALS

Types of materials. *Ancient materials.* Historically libraries have depended on what materials were available to build collections. The evolution of libraries in antiquity involved the search for a material durable enough to survive as a permanent record and relatively easy to use. Clay and stone provided permanence, but inscribing the records required considerable labour. Palm leaves, bamboo strips, and papyrus offered a flat surface that more readily accepted handwriting, and it was said that parchment came into use in Asia Minor after the export of papyrus from Egypt was banned. In AD 105 the invention of paper was announced by Ts'ai Lun (Cai Lun) to the Chinese emperor Ho-ti, and the British Museum has a paper fragment dated about 137. The use of paper spread slowly, however, and most of the oldest surviving manuscripts are of other materials, particularly vellum (fine-grained lambskin, kidskin, or calfskin).

Samples of ancient writing are rare and therefore are highly valued, and national and other scholarly libraries collect and preserve them as part of their responsibility to the preservation of history and the advancement of learning. Most universities have collections of rare books. Eton College, for example, has a fine collection of incunabula, some of which were purchased when they were first printed. A Gutenberg Bible is one of its finest examples. Some, such as the Duke Humphrey Library in the Bodleian at Oxford and the Beinecke Library at Yale University, contain collections of manuscripts, and wealthy private collectors have established world-famous institutions such as the Henry E. Huntington Library in San Marino, Calif., the Folger Shakespeare Library in Washington, D.C., and the Cotton and Harley collections in the British Library Reference Division.

Photographs. The invention of photography in the 19th century made possible a new kind of record, and collections of photographs are popular, particularly in public libraries with an interest in local history. Specialized photo libraries, such as the BBC Hulton Picture Library, are regularly used to provide illustrative material for film and television programs.

Uses of roll film and microfiche

For the general use of libraries, however, microphotography has played a much more important role. Many leading newspapers and periodicals have reproduced their entire sets of back issues on roll film, which offers a considerable saving of space and makes it feasible for even a small library to house an entire set. The disadvantage of roll film is that the user must start searching the roll film from the beginning of the reel, no matter where the relevant pages may be on the reel. A considerable advance was achieved by the invention of the transparent Microcard, or microfiche. This is a piece of film cut to a specified size and shape usually approximating a library catalog card but available in more than one size (although the most favoured size is 5 by 3 inches [8 by 13 centimetres]). The microfiche offers the advantage of random access; that is, instead of starting at the beginning, the user can bring any section of the microfiche directly into view on the screen. Microfiche also are more convenient to store and handle, and they have become very popular for the production of catalogs and bibliographies as well as for reproduction of texts.

Audiovisual materials. There are various forms of audiovisual media. The most common in libraries is the audio recording on disc or tape, and most libraries, especially public and school libraries, have built up extensive collections of nonbook materials, from the recordings of symphony orchestras on long-playing records or compact discs to tape-recorded oral history interviews. Videotape loans are also available from many libraries. Cooperation among school and public libraries in this field has made a considerable contribution to local history. The importance attached to these media, particularly in schools, is indicated by the use of the name *resource centre* for what was formerly called the school library. When teachers are eager to use audiovisual materials, they are often also eager to create materials, and this enthusiasm can enable school librarians to build up a strong and productive relationship between the library and the teaching program.

Magnetic materials. Newer articles requiring library storage include the machine-readable magnetic tape and disc. These need such specialized treatment, for the safeguarding of their contents from accidental erasure, that most computer centres employ their own specialist librarian. Like roll film, magnetic tapes and discs do not readily yield information about their contents and therefore require particular care in labeling and indexing as well as equipment and programming to permit retrieval.

Indexes and abstracts

Access to materials. Two types of documents, indexes and abstracts, contain catalogs and bibliographies of original materials. Indexes include any of countless bibliographies of currently published material, usually of articles in periodicals. Sometimes libraries have taken the initiative to create these finding aids for journal articles. For example, the U.S. National Library of Medicine has produced *Index Medicus,* a monthly listing of current articles from some 3,500 biomedical journals throughout the world. In other cases, scientific societies have taken the initiative. Early in the 20th century, the American Chemical Society began to prepare indexes and abstracts to help chemists obtain information about the literature in their field, and the Institute of Physics in the United Kingdom took a similar responsibility for physics. The long series of indexes published by the H.W. Wilson Company of New York City and covering many different fields is well known and widely used in other countries, though their coverage is mainly limited to American publications. This national focus has resulted in similar efforts in other countries, such as the *Current Technology Index* (British) and the *British Education Index.*

Since the 1960s many indexes and abstracts have become available electronically. In subject areas having a low level of user demand, libraries can obtain information about journal articles by connecting their computers to those of database vendors. In areas of higher demand, libraries can buy the electronic form of indexes and abstracts on CD-ROM (Compact Disc–Read-Only Memory). These small discs, which are identical in appearance to those used to record music, can replace several volumes of an index or abstracting service. In areas of the highest user demand, libraries can purchase the electronic indexes and abstracts on magnetic discs or tape and enter this information about journal and magazine articles into the same computer that contains the library catalog information. In this way, the OPAC (on-line public access catalog) can be extended to include information about both books and journal articles. One major advantage of electronic indexes and abstracts is that they enable users to search for information using words from titles and abstracts in addition to the author and subject access provided in print versions. This expands significantly the number of paths by which people can find information.

Abstracts, which are summaries of documents that indicate contents as well as authors and titles, have a history that dates to at least 1682, when the *Weekly Memorials for the Ingenious* was published in London. The *British Librarian* of 1737 published abstracts of well-known and useful books and claimed to cover all the sciences "in a manner never before attempted."

TECHNICAL SERVICES

One of two major functions of libraries, technical services include processes for acquiring, arranging, indexing, and storing the collection.

Acquisition and supply. *Criteria for selection.* The output of published materials, in all forms, is so vast that no single library, not even the largest, can hope to acquire everything; even in relatively specialized fields, some selection has become necessary, and most libraries have an explicit selection policy. The basic principles of selection vary little among different types of libraries, inasmuch as they derive directly from the known interests of the users. Practice is another matter and varies according to the types of user. A national library aims to hold at least one copy of all the publications of its own country and to have a good representation of foreign works, many of which may be obtained through exchange agreements with other national libraries. University, college, and school libraries relate their choice of acquisitions to the programs of teaching and research in their institutions; the academic level of the material naturally varies according to the level of the student population. An elementary school will hold a good selection of books written for children, but a university will tend not to. Many university libraries try to maintain a relatively complete coverage of the reports issued by government and other research establishments. Some universities are designated as repositories for the reports issued by intergovernmental agencies, such as the United Nations, the International Atomic Energy Agency, and the European Union.

An important aspect of selection is learning about new publications that would enhance the library. Various surveys have been made of the ways in which specialists gain new information about their fields of work, and the most popular usually turns out to be informal discussions with colleagues. But this is by nature a haphazard process, and most countries now have, or aim to have, a national bibliography based on the acquisitions of the national library. The *British National Bibliography,* begun in 1950 at the British Museum, is a leading example: it is published weekly, with regular cumulations for easy access over long periods. It is a tool for subject inquiry searches as well as for current selection.

The International Federation of Library Associations and Institutions has established a program to increase the range and number of such bibliographic tools. The program, called Universal Bibliographic Control and International MARC, aims to encourage national libraries, or groups of libraries, to institute methods of recording their national publications in a standard format and, wherever possible, of entering them into computer files. This program is accompanied by two additional programs, the Universal Availability of Publication and Universal Dataflow and Telecommunications, which aim to provide the necessary follow-up service of document delivery.

Other aids to the selection of material for acquisition are legion. Many libraries join professional societies and institutions to obtain their publications, which usually contain lists and reviews of new work relevant to their subjects. Leading journals, such as *The Times Literary Supplement, The New York Review of Books, Nature,* and *Science,* contain reviews by experts, advertisements for new and forthcoming publications, and review articles covering important new books in special fields.

Acquisition systems. The development of electronic means of document delivery is unlikely to supplant the more traditional sources of supply, the publishing and bookselling trades. Some companies combine the two functions. Purchases by libraries have traditionally generated much of the revenue of local bookshops, but firms operating as specialist library suppliers are able to offer many auxiliary services, such as attaching plastic covers and inserting labels of ownership, because they deal in large-scale bulk supply and can afford to maintain machines for such processes.

The acquisition systems described above are found mostly in countries with long-established traditions of reading, research, libraries, and book trade. Far greater difficulties confront the library services in the developing countries, particularly in Africa and Asia. Even in India and China, with their long history of using books, a steady and satisfactory progress is hindered by shortages of finance, materials, and trained staff. Some universities in these nations have large libraries and receive grants that enable them to acquire foreign as well as national publications, but they often meet with delays caused by administrative procedures, shortage of foreign currencies, and problems of language in the postal services. In most African countries, growth of a national literature is hampered by the centuries-old prevalence of an oral tradition and by the cost of importing even such basic materials as paper.

Many countries in eastern Europe as well as in the Third World look to exchanges as a means of obtaining materials. Some governments allow libraries to exchange duplicate copies of national publications, as a recognized method of compensation without payment in foreign currency. The practice does present certain administrative problems, but it is a useful means of encouraging the international flow of publications as well as of giving practical help in collection building to libraries in countries with limited resources.

Cataloging. However careful and scholarly the methods used in building a collection, without expert guidance to its access and use, the collection remains difficult to approach. Cataloging and classification, well-tried disciplines often combined under the general heading of "indexing," provide the needed guidance. Both techniques have been in use as long as libraries have existed, and their value in the so-called information age has been enhanced by the use of computers.

Function of the catalog

The function of the catalog is to identify all the items in a collection and to group like items together. All the great libraries of the ancient world seem to have had lists and inventories, whether kept on clay, stone, papyrus, parchment, palm leaves, or bamboo strips. Examples may be found in museums throughout the world.

Cataloging by author and subject. For many centuries the feature that gave a work its unique identity was the name of the writer, and users of the library were expected to know the names of the authors whose works they wished to consult. This system was eventually supplemented by the development of a subject catalog.

Many factors have contributed to the rise in importance of the subject approach to information. From the earliest times, librarians recognized that readers would be greatly helped if the catalog entries were arranged in groups of related subjects. General classifications of knowledge such as those of Aristotle and Porphyry, scholarly curricula such as the trivium and quadrivium (expounded by Julius Caesar's librarian Marcus Terentius Varro and persisting as a major influence on Western education), and practical considerations, such as the governmental needs of emperors and priests, all have formed the basis for the arrangement of subject catalogs. Early in the 7th century the scholar Wei Cheng wrote the bibliographic section of the official *Sui Dynasty History,* dividing the books into four categories: Confucian classics, historical records, philosophical writings, and miscellaneous works.

Research teams' effect on cataloging

Since the late 19th century far more attention has been paid to cataloging the subject contents of books as well as the names of their authors. Most of the impetus for this change came from science and technology, where the practice of working in teams in research institutions largely superseded the practice of single individuals, such as Charles Darwin, working for years to complete their research and then publishing the results as a book. The proliferation of specialist journals that publish short papers charting the progress of teamwork has meant that the names of single authors have become somewhat less important as tools for identifying works in libraries. Catalogs that list the subjects of research are more useful to specialists in related fields around the world, who may not know researchers by name but wish to have access to their work.

Catalog systems. Despite a steady, if slow, trend toward standardization, various forms of catalog continue to exist. Sets of entries generally are arranged in one of three catalog systems. The first is the dictionary catalog, in which author, title, subject, and any other entries are filed in a single alphabetical sequence. This form is popular in the United States and in public libraries generally and probably presents the least amount of difficulty for the general or casual reader. The second is the divided catalog, still in alphabetical sequence but with subject entries in a separate file. This form has increased in popularity, and many libraries have divided their former dictionary catalogs, recognizing the growing value of the subject approach. The third is the classed, or classified, catalog, which is more popular in Britain and continental Europe and in some developing countries whose librarians trained there. In the classed catalog, as its name suggests, all the entries are filed in the sequence of a classification scheme—that is, in a systematic order of subjects—but separate alphabetical files link names of authors and of subjects to the notation symbols of the classification scheme used in the main file. The chief advantage of a classed catalog is that the entries are related subjects grouped together in the file; thus, a subject search can be made much more simply than in a catalog based on the alphabet. In addition, when different languages are used, the sequence of entries in a classed catalog does not alter, as is the case with the dictionary.

Vehicles for catalogs. The types of catalog differ on the basis of the information provided in the entries, but the actual physical form may also vary. Originally, catalogs took the same form as the books they listed; being made of the same material, the catalog was an extra item of the collection itself. The earliest catalogs of the great national and scholarly libraries were in book form, with handwritten entries and spaces for new additions. The main problem of the book-type catalog, of course, was the insertion of entries for new acquisitions. Most plans, like that of the Bibliothèque Nationale in Paris, made no attempt to add to the printed file and instead placed vain hopes in the prospect of new editions.

Book catalogs

Card catalogs

The initial solution to this problem was the creation of a card catalog, each entry having its own card and each card containing only one entry. In principle, such catalogs can grow in size indefinitely; any new entry can be filed between any two existing entries. Thus the catalog offers the opportunity to have a completely up-to-date file: an entry can be made in the catalog immediately after a book has been purchased.

In 1901 the Library of Congress began offering copies of its own catalog cards for sale to other libraries. For many years this proved of inestimable value, particularly to small libraries unable to afford skilled catalogers. The service was also intended to serve as a central cataloging agency. Many eminent librarians, in conferences and in published papers, had lamented what they argued was wasteful duplication of effort involved in the separate cataloging of the same books in many different libraries. They proposed that a central agency undertake the task and make the results generally available, so that any library could use the central catalog thus produced to complete its own highly professional catalog. In the 1950s the British National Bibliography also began to produce cards from the entries in its weekly lists.

These and similar schemes in other countries in Europe achieved a certain success but for various reasons could not be said to have provided the ultimate solution. The advent of the computerized catalog, however, offers a more practicable approach because the storage capacity and the operating speed of even small machines overcome the main drawbacks to card services: delays in production and the labour of filing the cards when they arrive. Computer technology makes it possible for the details of any document to be entered into a file at any point and then to be transmitted to a central data file from which other libraries can obtain details by means of telecommunications links. The process is demonstrated by the revised Machine-Readable Cataloging Project, known since its revision in 1968 as MARC II. Library users find no difficulty in consulting such on-line catalogs, and many prefer them to the more cumbersome, if more familiar, form of cards in drawers. Not only do they enable library patrons to search for books containing a particular word or combination of works in the title, but the patron also can narrow a subject search by finding books that are on two or more topics. The information from the catalog can be alphabetized or put in order by year of publication and printed out as a book list.

Computerized catalogs

The system that accommodates this type of search is known as OPAC (on-line public access computer). Further development of this system has made it possible to integrate other library records with the OPAC, so that patrons can reserve materials that are still on order and can determine if items in the library's collection are already on loan. The OPAC has been expanded in many libraries to include information about journal articles and sometimes about the community served by the library. The job of cataloger, once highly dedicated to the task of describing library materials for the production of a catalog, is now focused on producing an information retrieval tool that will be of general use to the library's patrons.

Catalog standardization. The ideal of centralized cataloging led to increased interest in standardized forms of entry. As libraries grew larger after the Renaissance, it became necessary to devise some form of standard to ensure consistency among several catalogs. Perhaps the most famous, the British Museum Rules, was inspired by Sir Anthony Panizzi and has influenced all succeeding codes, though most of them have departed considerably from the original. The Vatican Rules and the Prussian Instructions have both been subject to commissions for revision, but certainly the most influential code is the Anglo-American *Catalog Rules; Author and Title Entries,* first published in 1908 and revised in 1967. A further revision was published in 1978 as *Anglo-American Cataloguing Rules,* second edition; it is commonly referred to as *AACR2.*

Anglo-American *Catalog Rules*

Many other discussions, revisions, and simplifications have taken place since the mid-20th century. Short versions of the major codes were published for small libraries in certain countries; the public instruction ministry in Italy issued new rules; a French commission on cataloging issued standards for anonymous works; and in the former Soviet Union, proposals were published for standardizing the transcription of Chinese names into Cyrillic script.

All these codes dealt with the use and entering of a consistent form of the names of authors, including Anonymous in the case of anonymous works, and sometimes, as in *AACR2,* with titles. A separate set of codes for subject cataloging emerged mainly in the United States. These subject headings took the form of lists of subject headings, the three best known being the list compiled by Minnie E. Sears, the *Library of Congress Subject Headings,* and the Medical Subject Headings (MeSH) of the U.S. National Library of Medicine.

Thesauri. A new use of the term *thesaurus,* now widespread, dates from the early 1950s in the work of H.P. Luhn, at International Business Machines Corporation (IBM), who was searching for a computer process that could create a list of authorized terms for the indexing of scientific literature. The list was to include a structure of cross-references between families of notions, in the manner of P.M. Roget's *Thesaurus of English Words and Phrases* (1852) and similar to the structure of faceted classification schemes. A major thesaurus, and one of the earliest, is the *Thesaurofacet* (1969), a list of engineering terms in great detail designed by Jean Aitchison for the English Electric Company. The thesaurus has proved very useful both for indexing and for searching in machine systems. It is especially helpful in such areas as medicine, aerospace, and other scientific and technical fields.

Thesauri depend upon the concept of controlled vocabulary, subject headings organized into lists that help users locate the appropriate heading for their topic of interest and find related terms used for narrower or broader topics. One of the functions of controlled vocabulary is to select from what may be a large group of synonyms the one term that most accurately describes a topic. When libraries use that one term consistently, their users will know where to look to find materials on any topic of interest. Books on travel in Britain, for example, might be described as travel, tourism, or sightseeing in Britain, Great Britain, or the United Kingdom. By selecting one expression of this topic—for example, "Great Britain — Description and Travel"—a library ensures that all of the books on the topic will be grouped under one subject heading in the catalog.

Controlled vocabulary

The disadvantage of controlled vocabulary systems such as subject headings lists is that they are slow to evolve. New topics and new ways of thinking and talking about topics continually evolve, and, although there are ongoing efforts to keep controlled vocabulary lists as up-to-date as possible, they inevitably lag behind ordinary language usage. As a result, some topics may be difficult to find using the subject heading approach. For this reason, many library systems, particularly those that use computerized information retrieval techniques, combine controlled vocabulary such as subject headings with free vocabulary:

descriptions of the topic of books or other materials using ordinary language. In these systems, subject headings are supplemented with topic descriptions such as abstracts and summaries, which have no restrictions on descriptive vocabulary.

Classification. While catalogs aim to identify and list items in a collection, schemes of classification have a more general application in arranging documents in a sequence that will make sense and be helpful to the user. Because they display subjects, and not documents, they can be used in several libraries, and some indeed have found applications in many different countries. Like schemes for grouping entries in catalogs, classifications—whether of knowledge based on philosophical principles, of the subject faculties of universities, or of the pragmatic grouping of books on shelves—have formed the basis of many individual systems.

The Dewey Decimal system. The best known of all schemes for the classification of documents in libraries is the Dewey Decimal Classification, devised by Melvil Dewey in 1873 and published in 1876. Apart from being the first modern classification scheme for libraries, the Dewey system embodies two of Dewey's many contributions to the theory and practice of librarianship. First, he recognized that a systematic arrangement of books on shelves should make sense to the users; his scheme therefore reflected the dominant pattern of current thinking, exemplified by the "classificatory sciences." And second, he used decimals as notation symbols, which illustrated the way in which subjects were divided hierarchically, from main classes to specific topics. An example from the schedule for chemistry shows how numbers are subdivided:

540	chemistry and allied sciences
541	physical and theoretical chemistry
541.2	theoretical chemistry
541.3	physical chemistry
541.34	solutions
541.35	photochemistry
542	laboratories, apparatus, equipment

Another feature of the Dewey system is the mnemonics used for certain types of subdivisions. Thus, many subjects can be subdivided geographically by the use of the historical-geographic number as decimals:

900	general geography and history
970	history of North America
973	history of the United States

Combining with the art schedule, the number for history of art in the United States is obtained:

700	the arts
709	history of art
709.73	history of art in the United States

The Dewey Decimal system and the Library of Congress system, mentioned below, are the classification schemes most frequently used in North American libraries.

The Universal Decimal system. The Universal Decimal Classification, published in 1905 and preferred by scientific and technical libraries, was an immediate offspring of the Dewey system. Paul Otlet and Henri-Marie Lafontaine adapted the Dewey system as the basis for a much more detailed scheme suitable for use in a vast card index of books and periodical articles in classified order—a universal bibliography of recorded knowledge. While retaining the basic generic hierarchies, the Universal Decimal Classification makes far greater use of the technique of synthesis, by providing a series of auxiliary tables for aspects of subjects likely to appear in several parts of the main schedules. These tables are indicated by the use of symbols such as punctuation marks. The colon sign (:) indicates a relationship between any parts and is the most commonly used sign. The numeral 669.1 being the notation for iron and steel and 546.22 for sulfur, the compound subject can be indicated by the notation 669.1:546.22, sulfur in iron and steel.

Like the Dewey Decimal Classification, the Universal Decimal Classification has been translated into many languages, and it is in use in many European and Asian libraries. European libraries, in particular, have emphasized classification systems over subject heading systems, basing their subject catalogs on the classification system, with an alphabetic index to class numbers. The revision of the Universal Decimal Classification has become a responsibility of the International Federation for Information and Documentation (Fédération Internationale d'Information et de Documentation; FID).

The Library of Congress system. At the turn of the 20th century Herbert Putnam, the Librarian of Congress, decided to reclassify the library but rejected the Dewey system. His staff adopted a more pragmatic approach based entirely on the way in which the books were arranged in their subjects on the shelves. They also rejected the decimal notation, preferring a purely ordinal system combining letters and numbers, leaving blank spaces where they expected new subjects to develop. (Not all of their expectations have proved correct.) American libraries and some scholarly libraries elsewhere have found the scheme attractive for its depth of detail, inasmuch as it is based on a very large library. An additional advantage is that Library of Congress notations appear on the library's catalog cards and on computer tapes produced by the MARC project. It uses both alphabet letters and numbers for its classification codes.

The Bliss system. Although not widely used, the bibliographic classification system invented by Henry E. Bliss of the College of the City of New York (published in 1935 as *A System of Bibliographic Classification*) has made important contributions to the theory of classification, particularly in Bliss's acute perception of the role of synthesis and his insistence that a library scheme should reflect the organization of knowledge and the system of the sciences. His systematic auxiliary schedules, designed to achieve what he called composite specification, carry the synthetic principle into every subject area and give a far higher degree of flexibility than does a purely enumerative scheme such as the Library of Congress system. The Bliss Classification Association, founded in the United Kingdom in 1967, promotes the use and development of the Bliss classification scheme.

The Colon system. Perhaps the most important advance in classification theory has been made by the Indian librarian S.R. Ranganathan, whose extraordinary output of books and articles has left its mark on the entire range of studies from archival science to information science. He introduced the term *facet analysis* to denote the technique of dividing a complex subject into its several parts by relating them to a set of five fundamental categories of abstract notions, which he called personality, energy, matter, space, and time. He employed these in his Colon Classification system (1933), which is used in some Indian libraries but has found few followers elsewhere. Nevertheless, the ideas in the scheme, expounded in his *Colon Classification* (1933) and *Prolegomena to Library Classification* (1937), have influenced all later work in classification theory and practice, including subsequent editions of the Dewey, Universal, and Bliss systems.

The Marxist system. In China a scheme has been published that departs somewhat from the Anglo-American traditions and claims to reflect the structure of knowledge according to the principles of Marxist philosophy. It has an enumerative structure and may be distinguished by its detail of analysis of, and dependence on, the corpus of Marxist literature—a literature that, in Anglo-American schemes, usually occupies a relatively minor place.

Preservation. One model of library service is that collections are used so extensively that the materials disintegrate from heavy use. Librarians who espouse this ideal maintain that libraries are for use and that any materials not in active use should be removed from the collection, a process known as weeding. A competing model holds that much of the world's great literature is available to be read chiefly because libraries of the past preserved that literature. This model encourages the preservation of materials so that the intellectual and cultural heritage received and created by the current generation can be transmitted to future generations.

Preservation versus use

Libraries most active in the area of preservation are usually large research libraries, which have the largest collections and perhaps the greatest concern for future users. On the whole most libraries try to strike a balance between maximizing current use and preserving materials for future use. In recent decades, the move toward preservation of library materials has been given additional impetus by the discovery that much 19th- and 20th-century paper retains acid introduced in the manufacturing process and that this acid, combined with the effects of air pollution, is causing many books to disintegrate as they sit on library shelves.

Reformatting. In response to this problem, libraries have developed several preservation strategies. The most important method of preserving library materials has been reformatting. Brittle and crumbling books and photographs are preserved by photographing them on microfilm or, in some cases, by using scanners to create digital images on magnetic or optical disc. These less vulnerable formats can then be preserved in archives. Reformatting also enables the inclusion of library materials in other media, such as multimedia information services. The drawback of this process, of course, is the issue of technological obsolescence. If reformatting relies on technology that becomes obsolete, the preservation effort is seriously compromised. The task of reformatting all materials that used acidic paper, nitrate films, or other degradable materials is monumental, generally requiring cooperation between many libraries and a substantial infusion of government funds.

Deacidification. In certain cases, reformatting is not the best solution to the problem of disintegration. The original material may have intrinsic value as an artifact, or it may lose some of its information in the reformatting process. In such cases, paper materials are deacidified by one of a number of chemical processes, some of which can also strengthen paper that has already been weakened. Mass deacidification of paper is an increasingly important part of preservation.

Future-conscious manufacturing. The most sensible solution to the preservation of books and journals for future use is the adoption of nonacidic paper by publishers. Many paper plants have begun to make a nonacidic paper that, with good care, will last for centuries rather than decades. Use of this product for book production will obviate the crumbling away of centuries of intellectual and cultural activity.

USER SERVICES

The second of the two main functions of libraries is directed at actively exploiting the collection to satisfy the information needs of library users.

Circulation. Although many of the libraries in antiquity were accessible to the literate public, this was almost certainly for reference only. Some monastic libraries, however, are known to have allowed the monks to borrow books for study in their cells; the Rule of St. Benedict explicitly permitted this, and the librarian exacted penance from any monk unable to confirm that he had actually read his book. Some university libraries may have lent books to members of their faculties, but the notion of lending, or circulating, libraries did not become popular until the 18th century.

Introduction of circulation systems

The rapid development of public libraries in the 19th century led to the extension of the practice and to the introduction of various systems for the recording of loans. All the early systems depended on the use of one or more cards on which were recorded the name of the book, the name of the borrower, and the date on which the book should be returned. Many libraries now use a computerized circulation system that records information about both the user and the material in circulation.

Reference and retrieval. Open access to the shelves and the facility to borrow books mean that much of the use of a modern library is at the free choice of the reader; scholars and scientists continue to emphasize the value of browsing among the shelves of a well-arranged library. "Chance favours only the prepared mind," said Louis Pasteur, and serendipitous discoveries of useful information during the search for some other subject have become a familiar and welcome aspect of using a library or other information service.

The role of the librarian

In reference service, librarians have traditionally given personal help to readers in making the best use of collections to satisfy their information needs. The publication of printed catalogs and bibliographies, the accessibility of on-line catalogs and multimedia databases, and the organizing of interlibrary cooperation have widened the range of resources available to the individual reader. As a result, librarians increasingly are called upon to help users determine the most efficient tool to use in their research. In scholarly libraries, assistance to readers once was generally limited to explaining the layout of the library and the use of the catalog; in universities, members of the faculty would have been expected to know the literature of their subject better than any librarian. But in public libraries, and still more so in special libraries in the fields of science and technology, readers have long sought guidance about information on their subject as well as about the library. This process has been greatly extended by the enormous increases in research worldwide and in the quantity of information and publications available in many languages and by the excellence of the indexes, abstracts, bibliographies, and databases that help to control the documentation of this massive output.

The subject search is one of the areas on which advancing technology has had the greatest impact. In many libraries, a variety of computer-based information retrieval systems provide ready access to details about off-site as well as on-site materials. For example, one development in subject access increases the amount of information that is available within library catalogs by including details from the table of contents or from a book's index. Other systems, rather than just listing the abstract of a journal article, include the entire article. These full-text references eliminate the intermediate step characteristic of older systems in which users first performed an electronic search and then obtained the articles themselves in print or microfilm.

Reference services can be broadly divided into two main aspects, usually known as retrospective searching (or information retrieval) and current-awareness service (or selective dissemination of information). These terms indicate a specialization that has occurred in this core activity of libraries and that grew mainly out of the expansion of scientific and industrial research during and after World War I. Three factors strongly influenced this process. First, the increase in research and publication affected all types of libraries and brought with it a similar increase in subject specialization. Second, working scientists, accustomed to referring to reports in published papers, were content to leave the organizing of information searches to a colleague who knew and understood their work. And, third, the widespread application of scientific research in industry provided an extra stimulus to the division of labour because of the necessity for speedy application of results to gain commercial success in production.

Some information specialists have tried to draw a distinction between their reference work and the more general reference services of librarians, but in most countries there is close cooperation among all engaged in these professional activities. Most acknowledge a mutual interest and influence, while the range of duties allows, indeed requires, different emphases in different institutions.

All agree, however, in acknowledging the duty to assist users to find answers to inquiries and to carry out searches in existing literature. Such a service requires many qualities, personal and professional: a detailed knowledge of books, periodicals, and all other forms of record; an ability to search efficiently in catalogs, indexes, abstracts, and databases; and, above all, a sensitive understanding of each user's needs. Matching the terms used by a reader in posing a question to the terms used by authors, indexers, and catalogers may well constitute one of the subtlest of professional skills.

Retrospective searching. The outcome of a search can take many forms, from a short, factual statement that gives the needed information to a short list of relevant references or a full-scale bibliography. In a computer search the first request often reveals that the database contains hundreds or even thousands of "hits," or references relating to the topic requested. The number can be reduced by narrowing the subject, adding more specific details, or defining more precisely the information needed. When a reasonable number of hits has been reached, the computer can be instructed to display the details of a few references to show the reader whether or not the search has covered the right subject area. If it has, the set of references or abstracts may then be obtained in the form of a printout; if it has not, the search begins again using new terms for the request.

In the specialized information centre a professional researcher can conduct the search and provide a state-of-the-art review of the literature in narrative form instead of as a collection of references. The service represents a peak of efficiency on behalf of the client, who has neither the time nor the resources to make the same review. The value to industry and commerce has encouraged private individuals to become information brokers—*i.e.,* to provide these services as a commercial enterprise.

Current-awareness service. The purpose of a current-awareness service is to inform the users about new acquisitions in their libraries. Public libraries in particular have used display boards and shelves to draw attention to recent additions, and many libraries produce complete or selective lists for circulation to patrons. Some libraries have adopted a practice of selective dissemination of information (sometimes referred to as SDI), whereby librarians conduct regular searches of databases to find references to new articles or other materials that fit a particular patron's interest profile and forward the results of these searches to the patron.

Selective dissemination of information (SDI)

One development of the concept of SDI in the electronic environment is a computer program that scans computer bulletin boards, electronic mail messages, and similar networked information resources and selects items that meet a user profile. Such programs enable individual users to keep abreast of the large amount of information available through computer networks without having to sift through much material that may be of little interest or relevance to them.

Community information. *Library extension programs.* The growth of information services in special libraries, followed by college and university libraries, also has influenced public library practice in library extension programs and community information services. Extension programs are usually arranged in cooperation with local educational organizations, university extramural courses, parent-teacher associations, and so on. In developing countries with nascent publishing and book trades, public libraries can offer valuable assistance to local authors, particularly those writing in indigenous languages, by providing facilities for authors to give lectures, hold seminars, and develop their own skills in direct relation with their potential readers. In European, African, and American libraries, poets or writers in residence have appeared as a part of similar action to bring authors and readers together.

Community awareness programs. In North America, community needs for informal information are often met by the public library's community awareness service (or information and referral service), though practice is far from standardized. This community outreach program is an important feature in many mostly rural societies. The Jamaica Library Service, for example, has long made a practice of setting up a stall at farmers' markets to supply up-to-date books and pamphlets on agriculture. Public libraries in China regularly set up special links with local factories for the supply of technical literature and specialist advisory staff.

INTERLIBRARY RELATIONS

Library cooperation. *Interlibrary lending.* The publication of bibliographies and library catalogs heightened awareness that no library could afford to be self-sufficient, and this awareness in turn stimulated interest in various forms of interlibrary cooperation. Cooperation probably originated informally, with readers referring to union catalogs to locate libraries that contained the books they wanted. One of the earliest formal organizations began with the Central Library for Students, founded in London by Albert Mansbridge in 1916. This was transformed in 1930 into the National Central Library, which continued to act as a lending library but also formed the centre of a network of regional library bureaus. The bureaus were located in a major regional library and, with one exception, built up union catalogs of holdings in the local public libraries to facilitate interlibrary lending. The National Central Library encouraged other university and special libraries to participate. The National Central Library has since become part of the British Library Lending Division, which undertakes a major part of interlibrary lending both in the United Kingdom and internationally.

The progress of interlibrary lending, coupled with the great losses suffered by libraries in Europe and Asia during World War II, led to an interest in cooperative acquisition of new materials. In 1948 the British National Book Centre was set up at the National Central Library in London to gather unwanted duplicates and to distribute them to the libraries that had suffered losses. It proved to be of incalculable value and was soon followed by the United States Book Exchange; both distributed lists of wants and offers to their member libraries.

British National Book Centre

Cooperative acquisition and storage. An ambitious program for cooperative acquisition of foreign materials by American libraries was conceived in the Library of Congress in 1942. This was the Farmington Plan: it involved the recruitment of purchasing agents in many countries, whose task was to buy their countries' current publications and distribute them to American libraries according to a scheme of subject specialization. Many criticisms were leveled at the scheme, and as a blanket operation it inevitably acquired a certain amount of trivia; but many research libraries have benefited by the acquisition of materials that otherwise would have been difficult to obtain.

Pressure on library space spurred librarians to discuss means of cooperative storage. Perhaps the foremost example is the Center for Research Libraries (formerly the Midwest Interlibrary Center) in Chicago, which began in 1952 as a centre for deposit of duplicate and little-used materials from research libraries. With the aid of a special grant, the University of London established a depository library, at Royal Holloway College away from the centre of London, to which the colleges of the university can send materials for either cooperative or private storage. The British Library Lending Division also acts as a cooperative store; it receives unwanted items from any library and makes them generally available. Both of these libraries reserve the right to refuse items that they already have in cooperative storage.

Cooperative cataloging. A number of important organizations facilitating library cooperation have been established to store and retrieve catalog records. In the United States, a library cooperative in Ohio grew into the OCLC Online Computer Library Center, a not-for-profit company with a database of millions of catalog records to which libraries can purchase access. Other organizations that store catalog records for retrieval by participating libraries have a regional focus or serve only one type of library, such as research libraries. In Canada, two organizations developed for storing bibliographic data; one is in the National Library of Canada and the other started as the cataloging centre of the University of Toronto and evolved into a for-profit corporation selling bibliographic records and other computer services to libraries. In Britain, the British Library became the supplier of cooperative library cataloging to all British libraries. Other library networks have been built around common automation choices. Groups of libraries that use the same circulation system or OPAC frequently organize to exchange information, advice, and bibliographic records.

Associations and international organizations. The wide variety of interlibrary organizations illustrated above makes for a dynamic and flexible infrastructure supporting library cooperation. Many library networks evolve from one type of organization into another. New organizations come into existence and old ones cease to function. Against this pattern of change, library associations provide a steady influence in favour of library cooperation. These associations, found at national, state, and local levels, provide a forum for discussing and adopting standards that encourage the sharing of resources. Such standards include the framework for interlibrary lending, the international cataloging codes and standards, and communications standards that allow library computer systems to be linked to each other.

International Federation for Information and Documentation

The oldest organization in the library and information field is the International Federation for Information and Documentation (see above, *The Universal Decimal system*). It was founded in 1895 in Brussels as the Institut International de Bibliographie by Paul Otlet and Henri-Marie Lafontaine, as part of their plan to create an index of world literature on cards. The institute has many international committees, and some, especially those concerned with classification research and the constant revision of the Universal Decimal Classification, are very active. The International Federation of Library Associations and Institutions (IFLA; Fédération Internationale des Associations de Bibliothécaires et des Bibliothèques, or FIAB) was founded in 1927 and first met formally in Rome in 1928. The organization publishes the *IFLA Journal*.

The International Council on Archives (ICA) was established with the help of UNESCO in 1948, and the first International Congress of Archivists was held in Paris in 1950. Early and continuing interest has centred on the microfilming, conservation, and preservation of historical records and on the development of standards for archive descriptions.

All these associations have received considerable moral and financial support from UNESCO, the first General Conference of which took place in 1947. From its inception UNESCO has placed great importance on the encouragement of bibliography and libraries, public libraries in particular. (Part of its program was inherited from a League of Nations organization called the International Institute of Intellectual Cooperation, a principal concern of which was libraries.) UNESCO's support has led to public library development in a large number of countries, as well as many other library-related projects.

The technical committee of the International Organization for Standardization, another United Nations body, has helped to formulate and promulgate a number of standards on bibliographical formats, particularly those related to computer processing. (D.J.F./Ed.)

BIBLIOGRAPHY

The changing role of libraries. Many authors speculate on the future of libraries in an electronic age. MICHAEL BUCKLAND, *Redesigning Library Services: A Manifesto* (1992), discusses what libraries should do to hold on to their enduring values while adapting to technological change. WILLIAM F. BIRDSALL, *The Myth of the Electronic Library: Librarianship and Social Change in America* (1994), provides an informed and insightful consideration of the impact of electronic information services on the social, cultural, and political role of the library. CHARLES R. MCCLURE, WILLIAM E. MOEN, and JOE RYAN (eds.), *Libraries and the Internet/NREN: Perspectives, Issues, and Challenges* (1994); and CHARLES R. MCCLURE, JOHN CARLO BERTOT, and DOUGLAS L. ZWEIZIG, *Public Libraries and the Internet: Study Results, Policy Issues, and Recommendations* (1994), discuss the impact of the Internet on library functions. Drawing on literature from both librarianship and other disciplines, MICHAEL A. HARRIS and STAN A. HANNAH, *Into the Future: The Foundations of Library and Information Services in the Post-Industrial Era* (1993), examines how libraries have used and been shaped by the growth of an information economy. GEOFFREY NUNBERG, "The Places of Books in the Age of Electronic Reproduction," *Representations*, 42:13–37 (Spring 1993), examines the book as a physical object in contrast to the electronic text. LINDA SCOVILL, *Librarians and Publishers in the Scholarly Information Process: Transitions in the Electronic Age* (1995), summarizes discussions of a closed meeting held between librarians and publishers. (L.S.E.)

The history of libraries. Much library history can be found in three subject encyclopaedias: ALLEN KENT *et al.*, *Encyclopedia of Library and Information Science*, 35 vol. (1968–83), continued with supplemental volumes (1984–); ROBERT WEDGEWORTH (ed.), *World Encyclopedia of Library and Information Services*, 3rd ed. (1993); and WAYNE A. WIEGAND and DONALD G. DAVIS (eds.), *Encyclopedia of Library History* (1994). JOHN FEATHER, *A Dictionary of Book History* (1986), helps define the intersection between print culture and library history. LUCIANO CANFORA, *The Vanished Library* (1989; originally published in Italian, 1986), is a highly readable account of the ancient Library of Alexandria. KARL CHRIST, *The Handbook of Medieval Library History*, rev. by ANTON KERN, trans. from German and ed. by THEOPHIL M. OTTO (1984), is still useful.

Types of libraries. Helpful works on types of libraries include MOHAMED MAKKI SIBAI, *Mosque Libraries: An Historical Study* (1987); JAMES THOMPSON (ed.), *University Library History: An International Review* (1980); and JESSE H. SHERA, *Foundations of the American Public Library: The Origins of the Public Library Movement in New England, 1629–1855* (1949, reissued 1974). Model library histories of particular countries include ANDRÉ VERNET (ed.), *Histoire des bibliothèques françaises*, 4 vol. (1988–92); and A.K. OHDEDAR (ĀDITYA OHADEDĀRA), *The Growth of the Library in Modern India: 1498–1836* (1966).

National libraries are dealt with in MAURICE B. LINE and JOYCE LINE (eds.), *National Libraries* (1979), and *National Libraries 2* (1987); and in the beautifully illustrated ANTHONY HOBSON, *Great Libraries* (1970).

Academic libraries are the subject of STEPHEN E. ATKINS, *The Academic Library in the American University* (1991); and ALLEN B. VEANER, *Academic Librarianship in a Transformational Age: Program, Politics, and Personnel* (1990). A more detailed discussion of the problems of handling large academic library collections is found in CLARE JENKINS and MARY MORLEY (eds.), *Collection Management in Academic Libraries* (1991).

General treatments of public libraries include W.J. MURISON, *The Public Library: Its Origins, Purpose, and Significance*, 3rd ed. (1988); and VERNA L. PUNGITORE, *Public Librarianship: An Issues-Oriented Approach* (1989). More detailed accounts of services to adults and to children are contained in KATHLEEN M. HEIM and DANNY P. WALLACE (eds.), *Adult Services: An Enduring Focus for Public Libraries* (1990); and MAE BENNE, *Principles of Children's Services in Public Libraries* (1991).

Two books that treat the impact of special libraries on organizations are JOSÉ-MARIE GRIFFITHS and DONALD W. KING, *Special Libraries: Increasing the Information Edge* (1993); and JOANNE G. MARSHALL, *The Impact of the Special Library on Corporate Decision-Making* (1993).

School libraries are covered in some detail in AMERICAN ASSOCIATION OF SCHOOL LIBRARIANS and ASSOCIATION FOR EDUCATIONAL COMMUNICATIONS AND TECHNOLOGY, *Information Power: Guidelines for School Library Media Programs* (1988); BETTY J. MORRIS, JOHN T. GILLESPIE, and DIANA L. SPIRT, *Administering the School Library Media Center*, 3rd ed. (1992); and JAMES E. HERRING, *School Librarianship*, 2nd ed. (1988).

Readers wishing to find more detail on archives may consult JAMES GREGORY BRADSHER (ed.), *Managing Archives and Archival Institutions* (1988); and RICHARD J. COX, *Managing Institutional Archives: Foundational Principles and Practices* (1992).

The library operation. ROBERT D. STUEART and BARBARA B. MORAN, *Library and Information Center Management*, 4th ed. (1993), focuses on general topics. The problems of developing and managing library collections are dealt with in several texts: BEATRICE KOVACS, *The Decision-Making Process for Library Collections: Case Studies in Four Types of Libraries* (1990); G. EDWARD EVANS, *Developing Library and Information Center Collections*, 2nd ed. (1987); and G.E. GORMAN and B.R. HOWES, *Collection Development for Libraries* (1989). Library techniques for organizing information are covered in LOIS MAI CHAN, *Cataloging and Classification: An Introduction*, 2nd ed. (1994); and BOHDAN S. WYNAR and ARLENE G. TAYLOR, *Introduction to Cataloging and Classification*, 8th ed. (1992). Issues involved in developing and managing library services to users are thoroughly covered in two standard textbooks: WILLIAM A. KATZ, *Introduction to Reference Work*, 6th ed., 2 vol. (1992); and RICHARD E. BOPP and LINDA C. SMITH (eds.), *Reference and Information Services: An Introduction* (1991). A good introduction to issues and practices of preservation of library materials may be found in JOHN N. DEPEW, *A Library, Media, and Archival Preservation Handbook* (1991).

ANNE WOODSWORTH and THOMAS B. WALL, *Library Cooperation and Networks: A Basic Reader* (1991), is an introductory text discussing interlibrary cooperation. The development of library associations is chronicled in WILLEM R.H. KOOPS and JOACHIM WIEDER (eds.), *IFLA's First Fifty Years: Achievement and Challenge in International Librarianship* (1977); W.A. MUNFORD, *A History of the Library Association, 1877–1977* (1976); and WAYNE A. WIEGAND, *The Politics of an Emerging Profession: The American Library Association, 1876–1917* (1986). (Ed.)

Life

The profusion of life on Earth has been studied in great detail, and a number of general principles have been revealed. Foremost among them is the principle of evolution by natural selection—the stepwise adaptation of organisms to their environment with increasing precision by small random mutations, or changes, in their hereditary material—which is the feature that distinguishes living from non-living matter. This article treats first the varieties of definitions of life and then covers, in some detail, the similarities and differences among organisms on Earth. It deals with the problem of the origin of life on Earth and concludes with a consideration of the possibility of life beyond the Earth.

This article is divided into the following sections:

DEFINITIONS OF LIFE

What is known about life

A great deal is known about life. Anatomists and taxonomists have studied the forms and relations of more than a million separate species of plants and animals. Physiologists have investigated the gross functioning of organisms. Biochemists have probed the biological interactions of the organic molecules that make up life on our planet. Molecular biologists have uncovered the very molecules responsible for reproduction and for the passage of hereditary information from generation to generation, a subject that geneticists had previously studied without going to the molecular level. Ecologists have inquired into the relations between organisms and their environments, ethologists the behaviour of animals and plants, embryologists the development of complex organisms from a single cell, evolutionary biologists the emergence of organisms from pre-existing forms over geological time. Yet despite the enormous fund of information that each of these biological specialties has provided, it is a remarkable fact that no general agreement exists on what it is that is being studied. There is no generally accepted definition of life. In fact, there is a certain clearly discernible tendency for each biological specialty to define life in its own terms. The average person also tends to think of life in his own terms. For example, the man in the street, if asked about life on other planets, will often picture life of a distinctly human sort. Many individuals believe that insects are not animals, because by "animals" they mean "mammals." Man tends to define in terms of the familiar. But the fundamental truths may not be familiar. Of the following definitions, the first two are in terms familiar in everyday life; the next three are based on more abstract concepts and theoretical frameworks.

Physiological. For many years a physiological definition of life was popular. Life was defined as any system capable of performing a number of such functions as eating, metabolizing, excreting, breathing, moving, growing, reproducing, and being responsive to external stimuli. But many such properties are either present in machines that nobody is willing to call alive, or absent from organisms that everybody is willing to call alive. An automobile, for example, can be said to eat, metabolize, excrete, breathe, move, and be responsive to external stimuli. And a visitor from another planet, judging from the enormous numbers of automobiles on the Earth and the way in which cities and landscapes have been designed for the special benefit of motorcars, might well believe that automobiles are not only alive but are the dominant life form on the planet. Man, however, professes to know better. On the other hand, some bacteria do not breathe at all but instead live out their days by altering the oxidation state of sulfur.

Metabolic. The metabolic definition is still popular with many biologists. It describes a living system as an object with a definite boundary, continually exchanging some of its materials with its surroundings, but without altering its general properties, at least over some period of time. But again there are exceptions. There are seeds and spores that remain, so far as is known, perfectly dormant and totally without metabolic activity at low temperatures for hundreds, perhaps thousands, of years but that can revive perfectly well upon being subjected to more clement conditions. A flame, such as that of a candle in a closed room, will have a perfectly defined shape with fixed boundary and will be maintained by the combination of its organic waxes with molecular oxygen, producing carbon dioxide and water. A similar chemical reaction, incidentally, is fundamental to most animal life on Earth. Flames also have a well-known capacity for growth.

Biochemical. A biochemical or molecular biological definition sees living organisms as systems that contain reproducible hereditary information coded in nucleic acid molecules and that metabolize by controlling the rate of chemical reactions using proteinaceous catalysts known as enzymes (see BIOCHEMICAL COMPONENTS OF ORGANISMS: *Enzymes*). In many respects, this is more satisfying than the physiological or metabolic definitions of life. There are, however, even here, the hints of counterexamples. There seems to be some evidence that a virus-like agent called scrapie contains no nucleic acids at all, although it has been hypothesized that the nucleic acids of the host animal may nevertheless be involved in the reproduction of scrapie. Furthermore, a definition strictly in chemical terms seems peculiarly vulnerable. It implies that, were a person able to construct a system that had all the functional properties of life, it would still not be alive if it lacked the molecules that earthly biologists are fond of—and made of.

Genetic. All organisms on Earth, from the simplest cell to man himself, are machines of extraordinary powers, effortlessly performing complex transformations of organic molecules, exhibiting elaborate behaviour patterns, and indefinitely constructing from raw materials in the environment more or less identical copies of themselves. How could machines of such staggering complexity and such stunning beauty ever arise? The answer, for which today there is excellent scientific evidence, was first discerned by the evolutionist Charles Darwin in the years before the publication in 1859 of his epoch-making work, the *Origin of Species*. A modern rephrasing of his theory of natural selection goes something like this: Hereditary information is carried by large molecules known as genes, composed of nucleic acids. Different genes are responsible for the expression of different characteristics of the organism. During the reproduction of the organism the genes also reproduce, or replicate, passing the instructions for various characteristics on to the next generation. Occasionally, there are imperfections, called mutations, in gene replication. A mutation alters the instructions for a particular characteristic or characteristics. It also breeds true, in the sense that its capability for determining a given characteristic of the organism remains unimpaired for generations until the mutated gene is itself mutated. Some mutations, when expressed, will produce characteristics favourable for the organism; organisms with such favourable genes will reproduce preferentially over those without such genes. Most mutations, however, turn out to be deleterious and often lead to some impairment or to death of the organism. To illustrate, it is unlikely that one can improve the functioning of a finely crafted watch by dropping it from a tall building. The watch may run better, but this is highly improbable. Organisms are so much more finely crafted than the finest watch that any random change is even more likely to be deleterious. The accidental beneficial and inheritable change, however, does on occasion occur; it results in an organism better adapted to its environment. In this way organisms slowly evolve toward better adaptation, and, in most cases, toward greater complexity. This evolution occurs, however, only at enormous cost: man exists today, complex and reasonably well adapted, only because of billions of deaths of organisms slightly less adapted and somewhat less complex. In short, Darwin's theory of natural selection states that complex organisms developed, or evolved, through time because of replication, mutation, and replication of mutations. A genetic definition of life therefore would be: a system capable of evolution by natural selection.

Instructions for living systems

Genetic definition of life

This definition places great emphasis on the importance of replication. Indeed, in any organism enormous biological effort is directed toward replication, although it confers no obvious benefit on the replicating organism. Some organisms, many hybrids for example, do not replicate at all. But their individual cells do. It is also true that life defined in this way does not rule out synthetic duplication. It should be possible to construct a machine that is capable of producing identical copies of itself from preformed building blocks littering the landscape but that arranges its descendants in a slightly different manner if there is a random change in its instructions. Such a machine would, of course, replicate its instructions as well. But the fact that such a machine would satisfy the genetic definition of life is not an argument against such a definition; in fact, if the building blocks were simple enough, such a machine would have the capability of evolving into very complex systems that would probably have all the other properties attributed to living systems. The genetic definition has the additional advantage of being expressed purely in functional terms: it does not depend on any particular choice of constituent molecules. The improbability of contemporary organisms—dealt with more fully below—is so great that these organisms could not possibly have arisen by purely random processes and without historical continuity. Fundamental to the genetic definition of life then is the belief that a certain level of complexity cannot be achieved without natural selection.

Thermodynamic. Thermodynamics distinguishes between open and closed systems. A closed system is isolated from the rest of the environment and exchanges neither light, heat, nor matter with its surroundings. An open system is one in which such exchanges do occur. The second law of thermodynamics states that, in a closed system, no processes can occur that increase the net order (or decrease the net entropy) of the system (see THERMODYNAMICS). Thus the universe taken as a whole is steadily moving toward a state of complete randomness, lacking any order, pattern, or beauty. This fate has been known since the 19th century as the heat death of the universe. Yet living organisms are manifestly ordered and at first sight seem to represent a contradiction to the second law of thermodynamics. Living systems might then be defined as localized regions where there is a continuous increase in order. Living systems, however, are not really in contradiction to the second law. They increase their order at the expense of a larger decrease in order of the universe outside. Living systems are not closed but rather open. Most life on Earth, for example, is dependent on the flow of sunlight, which is utilized by plants to construct complex molecules from simpler ones. But the order that results here on Earth is more than compensated by the decrease in order on the sun, through the thermonuclear processes responsible for the sun's radiation.

Life as an ordered system

Some scientists argue on grounds of quite general open-system thermodynamics that the order of a system increases as energy flows through it, and moreover that this occurs through the development of cycles. A simple biological cycle on the Earth is the carbon cycle. Carbon from atmospheric carbon dioxide is incorporated by plants and converted into carbohydrates through the process of photosynthesis. These carbohydrates are ultimately oxidized by both plants and animals to extract useful energy locked in their chemical bonds. In the oxidation of carbohydrates, carbon dioxide is returned to the atmosphere, completing the cycle. It has been shown that similar cycles develop spontaneously and in the absence of life by the flow of energy through a chemical system. In this view, biological cycles are merely an exploitation by living systems of those thermodynamic cycles that pre-exist in the absence of life. It is not known whether open-system thermodynamic processes in the absence of replication are capable of leading to the sorts of complexity that characterize biological systems. It is clear, however, that the complexity of life on Earth has arisen through replication, although thermodynamically favoured pathways have certainly been used.

The existence of diverse definitions of life surely means that life is something complicated. A fundamental understanding of biological systems has existed since the second half of the 19th century. But the number and diversity of definitions suggest something else as well. As detailed below, all the organisms on the Earth are extremely closely related, despite superficial differences. The fundamental ground pattern, both in form and in matter, of all life on Earth is essentially identical. As will emerge below, this identity probably implies that all organisms on Earth are evolved from a single instance of the origin of life. It is difficult to generalize from a single example, and in this respect the biologist is fundamentally handicapped as compared, say, to the chemist or physicist or geologist or meteorologist, who now can study aspects of his discipline beyond the Earth. If there is truly only one sort of life on Earth, then perspective is lacking in the most fundamental way.

LIFE ON EARTH

Mechanism and vitalism. Human beings are ambulatory collections of some 10^{14} cells. Human cells are in many fundamental respects similar to those that make up all the other animals and plants on the Earth. Each cell typically consists of a central, usually spherical, nucleus and an outer more heterogeneous region, termed the cytoplasm. The substance of nucleus and cytoplasm together has for many decades been called protoplasm. Use of this term implied that there was some special substance underlying living organisms. In the use of the word protoplasm there is occasionally an implication that life cannot be explained solely by physics and chemistry, that some mysterious "vital force" must be invoked. A living cell is a marvel of detailed and complex architecture. Seen through a microscope there is an appearance of almost frenetic activity. On a deeper level it is known that molecules are being synthesized at an enormous rate. Almost any enzyme catalyzes the synthesis of more than

100 other molecules per second. In 10 minutes, a sizable fraction of the total mass of a metabolizing bacterial cell has been synthesized. The information content of a simple cell has been estimated as around 10^{12} bits, comparable to about a hundred million pages of the *Encyclopædia Britannica.* Faced with all this or its equivalent, it is not surprising that early biologists felt despair at ever being able to understand the detailed workings of life.

A Stone Age man, confronted for the first time with a watch, might also deduce that there was some special watch substance in nature, or perhaps even a god of the watch. In ancient times, the most common of biological activities, such as the hatching of an egg or the blooming of a flower, were attributed to the intercession of a deity. After the epochal work of Sir Isaac Newton, when the motion of the planets and comets of the solar system was predictable to some very great precision and understood on the basis of an underlying principle, the idea developed that organisms were also nothing more than a particularly intricate kind of clockwork. But when early investigations failed to unveil the clockwork, a kind of ghostly mainspring was invented—the "vital force." This force was a rebellion from mechanistic biology, an explanation of all that mechanism could not explain or for which mechanism could not be found. It also appealed to those who felt debased by the implication that they were "nothing more" than a collection of atoms, that their urges and apparent free wills arose merely from the interaction of an enormously large number of molecules in a way that, although too complex to use predictably, was in principle determined.

"Vital force"

Not only is there no evidence for a vital force but the idea itself is hardly thought out; it is a sort of catchall concept, covering anything otherwise inexplicable. The alternative approach, that all organisms are made of atoms and nothing else, has proven especially useful and has led to a fundamental new understanding of biological systems. This situation does not imply, of course, that atoms cannot be put together in so complex a way that their collective behaviour is too difficult to understand in terms of the individual atoms; in this sense there may be particular laws of biology not readily derivable from the elementary interaction of atoms. But this is a very different thing from a vital force. Indeed, there is nothing debasing in the thought that a person is made of atoms alone; it means that one is intimately connected with the matter that comprises the inanimate universe. What a wonder that atoms can be put together in so complex a pattern as to produce human beings. Man is a tribute to the subtlety of matter. As the American anthropologist Loren Eiseley has written, ". . . if 'dead' matter has reared up this curious landscape of fiddling crickets, song sparrows, and wondering men, it must be plain even to the most devoted materialist that the matter of which he speaks contains amazing, if not dreadful powers. . . ." (*The Immense Journey,* Random House, New York, 1957.)

Nucleic acids. It is now known that many if not all of the fundamental properties of cells are a function of their nucleic acids, their proteins, and the interactions among these molecules. Within the nuclear regions of cells is a mélange of twisted and interwoven fine threads, the chromosomes. During cell division, in all but the simplest organisms, the chromosomes display an elegantly choreographed movement, separating so that each daughter cell of the original cell receives an equal complement of chromosomal material. This pattern of segregation corresponds in all details to the theoretically predicted pattern of segregation of the genetic material implied by the fundamental genetic laws (see GENETICS AND HEREDITY). The chromosomes are composed of nucleic acids and proteins in a combination called nucleoprotein. The nucleic acid stripped of its protein is known to carry genetic information and to regulate cellular metabolism; the protein in nucleoprotein undoubtedly plays some secondary, probably regulatory, role.

The role of nucleic acids

The specific carrier of the genetic information in higher organisms is a nucleic acid known as DNA, short for deoxyribonucleic acid. DNA is a double helix, two molecular coils wrapped around each other and chemically bound one to another by bonds connecting adjacent bases. Each helix has a backbone that consists of a long sequence of alternating sugars and phosphates. Attached to each sugar is a base. Each sugar-phosphate-base combination is called a nucleotide; a nucleic acid strand can be thought of as a sequence of nucleotides. There is a very significant one-to-one base pairing in the connection of adjacent helices, in the sense that once the sequence of bases along one helix is specified, the sequence along the other is also specified. The specificity of base pairing plays a key role in the replication of the DNA molecule, where each helix makes an identical copy of the other from molecular building blocks in the cell. These nucleic acid replication events are mediated by enzymes, and with the aid of enzymes have been produced in the laboratory.

Ribonucleic acid (RNA) differs from DNA in having a slightly different five-carbon sugar, and in replacing one of the four bases that make up DNA by a slightly different base. RNA does not appear to exist in a double-stranded form. Now DNA, RNA, and the enzymes have a curiously interconnected relation, which appears ubiquitous in all organisms on Earth today.

Commonalities among organisms on Earth. The genetic code was broken in the 1960s. It was found that three consecutive nucleotides code for one amino acid of a protein molecule; *e.g.,* an enzyme. By controlling the synthesis of enzymes, the nucleic acids control the functioning of the cell. Of the four different bases taken three at a time, there are $4^3 = 64$ possible combinations. The meaning of each of these combinations, or codons, is known. Most of them represent a particular amino acid. A few of them represent punctuation marks; for example, instructions to start or stop a synthesis. Some of the code is degenerate in the sense that more than one nucleotide triplet may specify a given amino acid. These interactions among nucleic acids and proteins seem absolutely central to living processes on Earth today. Not only are these processes apparently the same in all organisms on Earth but even the particular dictionary that is used for the transcription of nucleic acid information into protein information seems to be essentially the same in all organisms. Moreover, this code has various chemical advantages over other conceivable codes. The complexity, ubiquity, and advantages of these processes clearly argue that the present interactions among proteins and nucleic acids are themselves the product of a long evolutionary history. At the time of the origin of life this very complex replication and transcription apparatus could of course not have been in operation. A fundamental problem in the origin of life is the question of the origin and early evolution of the genetic code.

There are many other commonalities among organisms on Earth. For example, there is only one class of molecules that store energy for biological processes until the cell has use for it, and these molecules are all nucleotide phosphates. The most common example is ATP (adenosine triphosphate). For this very different function, a molecule identical to the building blocks of the nucleic acids is employed. There are metabolically important molecules—*e.g.,* molecules known as FAD (flavin adenine dinucleotide) and coenzyme A—which include subunits similar to the nucleotide phosphates. Porphyrins represent another category of ubiquitous molecules. Porphyrins are the chemical basis of hemoglobin, which carries oxygen molecules through the bloodstreams of animals; of chlorophyll, which is the fundamental molecule mediating photosynthesis in plants; and of the colours that many animals display. The left- or right-handedness of many biological molecules—discussed more fully below—runs identically through all organisms on Earth. In fact, of the billions of possible organic compounds, less than 1,500 are employed by contemporary life on Earth, and these 1,500 are constructed from less than 50 simple molecular building blocks. Similarly, organisms as diverse as paramecia and human sperm cells have little whiplike appendages called cilia or flagella used to propel themselves through their liquid environments. The cross-sectional structure of the cilia and flagella is almost always nine pairs of peripheral and one pair of internal fibres. There is no immediately obvious selective advantage of the 9 : 1 ratio.

Energy-storing molecules

These commonalities indicate that a few basic chemical and functional patterns are being used over and over again, reflecting the extremely close relations all organisms on Earth have to one another. Many biologists believe that these commonalities—particularly where no obvious selective advantage exists—imply that all organisms on the Earth are descendants of a single common ancestor. But it is possible that there are more subtle selective advantages. The issue may have to await the first detailed study of an extraterrestrial organism.

Chromosome possibilities

The number of possible ways of putting nucleotides together in a chromosome is enormous. The renowned geneticist H.J. Muller estimated that in a human chromosome there are about 4×10^9 base pairs. Each base pair position could be filled by any one of four possible bases; accordingly, the number of possible varieties of human chromosomes is $4^{4 \times 10^9}$ ($10^{2.4 \times 10^9}$), an inconceivably large number. By contrast, the number of elementary particles (electrons and protons) in the entire physical universe is only about 10^{80}. Thus a human being is an extraordinarily improbable object. Most of the $10^{2.4 \times 10^9}$ possible sequences of nucleotides would lead to complete biological malfunction. Our nucleotides work because natural selection, over a 4,000,000,000-year history of life, has destroyed enormous numbers of combinations that did not work. But there still may be combinations that work far better than any now present, and the future holds the promise that man will be able to assemble nucleotides in any desired sequence to produce whatever characteristics of human beings are thought desirable, an awesome and disquieting prospect.

Metabolism. The chemical bonds that make up living organisms have a certain probability of spontaneous breakage. Accordingly, mechanisms must exist to repair this damage, or to replace the broken molecules. In addition, the meticulous control that cells exercise over their internal activities requires the continued synthesis of new molecules. These processes of synthesis and breakdown of the organic molecules of the cell are collectively termed metabolism, and for synthesis to keep ahead of the thermodynamic tendencies toward breakdown, energy must be supplied to the living system.

Methods of acquiring energy

Organisms acquire this energy by two general methods. Some organisms are heterotrophs, acquiring their energy by the controlled breakdown of pre-existing organic molecules (food)—generally those supplied by other organisms. Human beings and most other animals are heterotrophs. Alternatively, organisms may be autotrophs, acquiring their useful free energy from some other source, either from the energy of sunlight, in which case the organism is called a photoautotroph, or from the controlled chemical reaction of inorganic materials, in which case the organism is known as a chemoautotroph. Organisms that use both modes are called photochemoautotrophs.

A green plant is a typical example of a photoautotroph. It uses sunlight to break water into oxygen and hydrogen. Hydrogen is then combined with carbon dioxide to produce such energy-rich organic molecules as ATP and carbohydrates, and the oxygen is released back into the atmosphere. Many animals, on the other hand, utilize the atmospheric oxygen to combine chemically with organic materials they have eaten and release carbon dioxide and water as waste products in extracting energy from the organic materials. This is an example of an ecological cycle in which a material (here carbon) is pumped through two different organisms.

More generally, such metabolic cycles—used by the organism to extract useful energy from the environment—can be described in terms of oxidation-reduction reactions. In the case of respiration, molecular oxygen accepts electrons from glucose or other sugars. The oxygen is said to be an electron acceptor (it has a great affinity for electrons), the glucose an electron donor. This is the prototype of oxidation-reduction reactions, but not all such reactions necessarily involve oxygen. Biological electron acceptors other than oxygen include nitrates, sulfates, carbonates, nitrogen, and methanol. Biological electron donors other than sugars include nitrogen, sulfides, methane, ammonia, and methanol. For acceptor-donor transformations to occur over any period of time, biological cycles are necessary. It is possible that, for geologically short periods of time, organisms have lived off a finite supply of material, but for any long-term continuance of life, a dynamic cycling of matter, involving at least two different varieties of organisms, is necessary. If there is life on other planets, a similar cycling must exist. A search for such molecular transformations is one method of detecting extraterrestrial life.

On the Earth, all such useful biological electron transfer reactions lead to the net production of one or more molecules of ATP. Two of the three phosphates of this molecule are held by "energy-rich" bonds sufficiently stable to survive for long periods of time in the cell, but not so strong that the cell cannot tap these bonds for energy when needed. ATP and very similar molecules, all of them having a base, a five-carbon sugar, and three phosphates, are, so far as is known, the general and unique energy currency of living systems on Earth.

Energy from metabolic processes

Metabolic processes do not occur in one step. The ordinary sugar, glucose, is not oxidized to carbon dioxide and water by living cells in the same way that occurs if a flame is applied to glucose in air. The resulting release of energy would be much too sudden, and concentrated in too small a volume, for such a process to be utilized safely by the cell. Instead, the glucose is broken down by a series of successive and coordinated steps, each mediated by a particular and specific enzyme. In almost all organisms that metabolize glucose, the sugar is first broken down in a set of anaerobic steps (that is, in the absence of oxygen). The total number of such steps is about 11. Some organisms are anaerobes; that is, they do not utilize molecular oxygen. In them the anaerobic steps are as far as the glucose metabolism is carried. Other organisms, including man, carry the oxidation of glucose further, gingerly combining glucose breakdown products with molecular oxygen. Such aerobic oxidation of glucose requires about 60 more enzymatically catalyzed steps. Another indication of the relative simplicity of the anaerobic breakdown of sugar is that all the enzymes used are free in solution in the cell; the aerobic steps use enzymes that are localized in specific regions of the cell.

Efficiency of aerobic oxidation

The complete aerobic breakdown of sugar to carbon dioxide and water is about 10 times more efficient than the breakdown accomplished by anaerobes; 10 times as many ATP molecules are produced. Similar themes and variations exist for the metabolism of other molecules (see METABOLISM).

The energy made available in this way to ATP is used in a variety of ways by the cell; for example, for motility. When an amoeba extends pseudopods, or a person walks, ATP molecules are being tapped for their energy-rich phosphate bonds. In addition, ATP molecules are used for the synthesis of molecules that the organism needs and does not have available. Among such molecules may be amino acids, the particular five-carbon sugars involved in nucleic acids, the nucleic acid bases, and so on. Each of these synthetic processes is again controlled and enzymatically mediated and may start from a variety of building blocks available to the organism, some simple, some more complex. For example, the amino acid L-leucine is produced from pyruvic acid, itself the product of the anaerobic breakdown of glucose. Synthesis of L-leucine from pyruvic acid involves eight enzyme-mediated steps and the addition of acetic acid and water.

These exquisitely interlocked and controlled metabolic steps are not usually performed in a diffuse manner all over the cell. Instead there is, at least in all higher organisms, a marvellously architectured cellular interior with particular specialized regions where particular chemical reactions are performed. Those oxidation-reduction reactions that involve molecular oxygen occur in an inclusion within the cytoplasm called the mitochondrion.

The work of the mitochondrion

The mitochondrion itself has an intricate substructure, and particular enzymes are thought to reside in particular sites within it; the molecule being metabolized may be passed on from one enzyme to another as through a conveyor belt in a factory. Similarly, photosynthesis occurs in a cytoplasmic inclusion called a chloroplast, which contains the chlorophyll and other pigments that absorb visible light, as well as the detailed enzymatic apparatus for the photosynthetic process. Chloroplasts and mitochondria, as well as other cytoplasmic inclusions at the base of flagella and cilia, all contain DNA. Moreover, this DNA has a somewhat different distribution of bases from that of the nucleus. It has been suggested that the cytoplasmic inclusions are the remnants of once free-living forms that, because of the favourable conditions found there, have taken up residence in the insides of other organisms.

Nucleic acids are known to pass from cell to cell and to perform their replication and coding functions efficiently in the new cell. In fact, viruses are essentially strands of nucleic acid, with a protein coat, that operate in just this way. It is also known that pieces of the genetic material from one cell may migrate into another cell of the same species and produce genetic and permanently heritable changes there. Alternatively, part of the virus nucleic acid may be permanently bound to the nuclear DNA of a host cell. It is likely that a virus is a degenerate form, now highly specialized to live off a specific host, of an organism once free-living and much more generally capable of performing a wide range of metabolic tasks. A virus must use the genetic transcription apparatus of its host cell. Many viruses do this extremely efficiently, turning a bacterium from a factory for making other bacteria into a factory for making viruses. In some cases it takes no more than 10 minutes for a bacterium infected by a single virus to produce a hundred new virus particles, which then burst forth from the host bacterium, destroying it. The line between benign or useful cytoplasmic inclusions and infective agents is not a very sharp one (see VIRUSES).

Eucaryotes and procaryotes. In the very simplest one-celled organisms one may distinguish between eucaryotic and procaryotic cells. Many familiar one-celled organisms, such as paramecia and amoebas, as well as the cells of all higher organisms including man, are eucaryotic. Such cells undergo mitosis, a fundamental sequence of events that occurs after DNA replication and that ensures that the DNA is precisely and equally distributed to the daughter cells. Eucaryotic cells have nucleoprotein in their nuclei. There is a membrane that separates the nucleus from the cytoplasm. Mitochondria are generally present in the cytoplasm, as is a very intricately convoluted structure, called the endoplasmic reticulum, that probably serves as the anchoring point for many cytoplasmic enzymes not contained in such inclusions as mitochondria or chloroplasts.

On the other hand there are procaryotic cells, which are most generally typified by the bacteria and the blue-green algae. In these cells nuclear division is nonmitotic, there is no nucleoprotein, and a nuclear membrane is absent. While eucaryotic cells may have more than one chromosome, procaryotic cells have one chromosome only, and that one is dispersed in the cytoplasm. Mitochondria, chloroplasts, and the endoplasmic reticulum are always absent. It is clear from this description that the procaryotes are in many respects more primitive than the eucaryotes. A basic unsolved evolutionary question concerns the evolution of procaryotes into eucaryotes.

The smallest organism

An interesting subject of biological speculation concerns what the smallest and simplest contemporary free-living organism might be. The smallest free-living cells now known are the pleuropneumonia-like organisms (PPLO). While an amoeba has a mass of 5×10^{-7} grams (1 gram = 0.035 ounce), a PPLO weighs 5×10^{-16} grams and is only about $^1/_{10}$ of a micrometre across. It can be seen only in the electron microscope. Such organisms grow very slowly. There may be smaller organisms that grow even more slowly, but they would be extremely difficult to detect. Even an organism of the size of PPLO has room for only about a hundred enzymes. A much smaller organism would have room for many fewer enzymes, and its ability to accomplish the functions that contemporary living systems must accomplish would be severely compromised. Were there, however, an environment in which all the necessary organic building blocks and such energy sources as ATP were provided "free," a functioning organism could be substantially smaller than PPLO. In fact the inside of a cell provides just such an environment; this is why infectious agents, such as viruses, can be substantially smaller than PPLO. But it must be emphasized that such agents are not free-living organisms.

Metazoa, embryology, and sex. The distinction between single-celled and many-celled organisms (in animals, between protozoa and metazoa) is far from a sharp one. An interesting illustration is the slime molds, which undergo an extraordinary sequence of events during their life cycle. The cycle begins with single cells, somewhat like amoebas, which swarm, or combine, into a slimy mass with many nuclei called a plasmodium. The plasmodium in turn forms a sluglike mass that is certainly a multicelled organism. The slug develops into a stalked, fruitlike sporangium, still multicellular. The sporangium produces spores with cellulose cell walls similar to those of plants. The spores in turn germinate into small cells bearing flagella. The flagella are lost and the life cycle is completed with the production of an amoeboid form.

Biology is replete with life cycles of comparable complexity. It is possible that the swarming of individual cells to form a plasmodium may in fact be an example of the events that led to the production of metazoa in the early history of the Earth. Such life cycles, while apparently very exotic, are shared by many organisms, including man, where a one-celled, free-swimming sperm stage is part of the life cycle.

The problem of development

The life cycle of slime molds, or humans, or any other multicellular organism, brings up a fundamental and still largely unsolved problem. These organisms develop from a single cell that has a single complement of the genetic material. These cells then divide, forming many identical cells. The very early embryology of man goes through stages with 2, 4, 8, 16, etc., cells. Since the genetic information is identical in each cell, how does it ever happen that cells become specialized, forming hair cells, teeth cells, liver cells, blood cells, or bone cells? How can any given cell "know" what sort of specialized cell it must become, since all cells contain identical nucleic acids? Possibly the answer to this question has to do with geometry. After the 16- or 32-cell stage, there is a distinct difference between a cell on the inside of the embryo, which is entirely surrounded by cells, and a cell on the outside of the embryo, which is not entirely surrounded by other cells. One of the earliest major steps in embryonic development is a distinction in function between interior cells (the endoderm) and exterior cells (the ectoderm). There are physical and chemical interactions among adjacent cells. Perhaps any cell then has the capability of becoming any specialized cell, but cells are, as a result of their external cellular environment, called upon to develop in different ways. Occasional embryonic anomalies or cysts occur in which, for example, hair or teeth develop in totally inappropriate portions of the body. Similarly, eyes have been caused to develop on the limbs of frogs. Such incidents demonstrate the capability of the "wrong" cells to produce particular cellular specializations (see GROWTH AND DEVELOPMENT).

Much of the beauty and diversity of contemporary life on Earth is due to sex. A totally asexual organism will be genetically identical to its (single) parent except for occasional mutations. The development of any major new adaptation would then require the acquisition of large numbers of appropriate mutations. Consider, for example, how the ability of an organism to metabolize a given molecule depends on the interaction of many enzymes, each produced by the transcription of the genetic information in hundreds of nucleotides, each nucleotide being the product of a single mutation. Thus, the chance development of any advantageous adaptation in an asexual organism requires the mutations to wait in line for a fortuitous juxtaposition.

Sex: an added dimension

Sex solves this problem in an elegant way. The genetic material of the parents is reassorted so that totally new combinations of genes are produced. In this way mutations acquired by any member of the population are rather quickly distributed to other members, and mutations arising in separate organisms can be combined. The likelihood of producing a useful sequence of mutations is thereby greatly enhanced. The advantages of sexual reproduction are so great that even many simple forms, such as bacteria or protozoa, which largely reproduce asexually, have occasional sexual encounters. While two sexes are clearly adequate for such a random assortment of genetic material, some organisms have developed more sexes: paramecia, for example, have somewhere between five and 10 sexes, defined in terms of the elaborate taboos about which organisms can combine their genetic material. In the process of genetic reassortment, some organisms make a very large number of attempts; for example, frogs lay millions of eggs at a time, and the number of sperm cells in a single human ejaculation is about 3×10^8 (see SEX AND SEXUALITY).

The varieties of organisms and environments. The environment of the Earth is heterogeneous. There are mountains, oceans, and deserts, extremes of temperature and humidity. In addition, there are diverse microenvironments: oxygen-depleted oceanic oozes, ammonia-rich soils, mineral deposits with a high radioactivity content, and so on. The environment of an organism also includes the other organisms in its surroundings. For each of these environmental situations there are corresponding ecological niches, and the variety of ecological niches populated on the Earth is quite remarkable. Furthermore, ecological niches can be filled independently several times. For example, quite analogous to the ordinary mammalian wolf is the marsupial wolf that lives in Australia; the two have striking similarities in physical appearance and in predation behaviour. As another example, the same streamlined shape for high-speed marine motion has evolved independently at least three times: in *Stenopterygius* and other Mesozoic reptiles; in the tuna, which are fish; and in the dolphins, which are mammals. This case of convergent evolution must arise from the fact that hydrodynamics admits a narrow range of solutions to the problem of high-speed marine motion by large animals. Similarly, the eye has independently evolved several times among animals on the Earth; apparently such a structure is the best solution to the problem of visual recording. In those cases where physics or chemistry establishes one most efficient solution to a given ecological problem, natural selection will often tend to reach the solution, but not always. Some adaptations of undoubted utility, such as tractor treads in swampy environments, have never been evolved by natural selection on the Earth.

Life in extreme environments

There is an extraordinarily wide range of ecological niches to which organisms have adapted through the operation of natural selection. The same basic fabric of life has been used to produce very diverse organisms. The alga *Cyanidium caldarium* can grow in concentrated solutions of hot sulfuric acid. Other bacteria, algae, and fungi can live in extremely acidic (pH of 0) or extremely alkaline (pH near 13) environments. Procaryotic bacteria live in pools at Yellowstone National Park at temperatures above 90° C (194° F), almost at the boiling point of water. Sulfate-reducing bacteria are reported to grow and reproduce at 104° C (219° F) under very high pressures. Many organisms employ organic or inorganic antifreezes to lower the freezing point of their internal liquids, so that they can live at several tens of degrees below 0° C (32° F). Some insects use dimethyl sulfoxide as an antifreeze. Other organisms live in briny pools in which dissolved salts lower the freezing point. For example, Don Juan Pond in Antarctica has about one molecule of calcium chloride for every two water molecules and does not freeze until −45° C (−49° F). It contains a possibly unique microflora that continues to metabolize at least down to −23° C (−9° F). Biological activity does not cease at the freezing point of water; in fact some enzymes are actually more active in ice than in water. Many single-celled organisms can be frozen indefinitely to extremely low temperatures—the temperature of liquid air for example—and then be thawed with no decrease in activity. The primary damage that freezing causes is apparently due to the unavailability of liquid water and to the expansion and contraction attendant to freezing and thawing. Some arthropods can be severely dehydrated and then revived simply by adding water. In the dehydrated state they can be brought to any temperature from close to absolute zero to above the boiling point of water without apparent damage. When encysted in response to dehydration, some such organisms seem indistinguishable from a weathered grain of sand.

The great majority of familiar organisms on the Earth, however, are much more sensitive to the temperature of their surroundings. Warm-blooded animals internally regulate their temperatures for this reason. A human being whose body temperature drops below 30° C (86° F) or rises above 40° C (104° F) soon dies. Organisms that inhabit cold climates have special insulating layers of fat and fur. Other organisms adapt to seasonal temperature changes by producing dormant forms, such as spores or eggs, to survive the low temperatures. In all cases dormancy appears to be accompanied by dehydration.

Importance of water

Since organisms are composed largely of water, the availability of water is clearly a limiting factor. Here also, however, remarkable adaptations exist. Certain microorganisms can live on the water adsorbed on a single crystal of salt. Other organisms, such as the kangaroo rat and the flour beetle, obtain no water at all in the liquid state, relying entirely on metabolic water; that is, on water released from chemical bonds through the metabolism of food. A variety of plants, including Spanish moss, live in environments where they have no contact with groundwater—for example, on telephone wires—apparently extracting water directly from the air, although such plants require a relatively high humidity. Plants that live in deserts and other very dry environments have evolved wide-spreading root systems that adsorb subsurface water from a great volume of adjacent soil.

Organisms have been found from the stratosphere to the ocean depths. Bacteria and fungal spores have been discovered near the base of the stratosphere by balloons, and searches for organisms at much greater altitudes (up to 100,000 feet) have been attempted with ambiguous results. Birds have been observed flying at altitudes as great as 27,000 feet, and jumping spiders have been found at 22,000 feet on Mt. Everest. At the opposite extreme, microorganisms, fish, and a variety of other metazoa have been recovered from the ocean depths down to thousands of feet, where the corresponding pressures are hundreds of times that at sea level. At these depths no light can penetrate and the organisms, some of which are quite large and include unique phosphorescent adaptations to the dark, ultimately live off particles of organic matter raining down from the upper reaches of the oceans.

There is a range of adaptations to the radiation environment of the Earth. Some microorganisms are readily killed by the small amount of solar ultraviolet light that filters through the Earth's atmosphere at wavelengths near 3,000 angstrom units (Å; 1 Å = one ten-billionth of a metre). On the other hand, the bacterium *Pseudomonas radiodurans* thrives in the large neutron flux at the cores of swimming-pool reactors, to the continuing annoyance of nuclear physicists. Organisms can avoid radiation by shielding. For example, some algae and some desert plants live under a superficial coating of soil or rocks that are more transparent to visible light than to ultraviolet light. In addition, organisms have active methods of undoing the damage produced by radiation. Some of these repair mechanisms work in the dark; others require visible light. The usual reason for the ultraviolet sensitivity of organisms is that their nucleic acids absorb ultraviolet light very effectively at a wavelength near 2,600 Å. Generally speaking, there is an upper limit to the amount of ionizing radiation (such as gamma rays, X-rays, electrons or protons) that an organism can receive without being killed: in the vicinity of 1,000,000 roentgens. Such a lethal dose applies only to extremely radiation-resistant microorganisms; mammals, for example, are killed by much lower doses because there is more that can go wrong with such complex organisms. A lethal dose of ionizing radiation for human beings is a few hundred roentgens applied to the whole body. A thermonuclear weapon dropped on a populated area may deliver, through direct radiation and fallout, doses of a few hundred roentgens or more to people within a radius of some tens of miles of the target. Much smaller doses can produce a variety of diseases and predominantly deleterious mutations in the hereditary material. Moreover, the effect of small doses is cumulative. But until very recently human beings have not lived in environments characterized by large doses of ionizing radiation (see RADIATION).

Size range of living things

The sizes of organisms on the Earth vary greatly. As discussed above, the smallest free-living organisms on the Earth, PPLO, are about 1,000 Å in diameter; a limitation on the size of the smallest free-living organism is its volume: it must contain all the molecules necessary for metabolism. A variety of influences place an upper limit to the size of organisms. One is the strength of biological materials. Galileo calculated in 1638 that a tree taller than roughly 300 feet (91.4 metres) would, when displaced slightly from the vertical (for example, by a breeze), buckle under its own weight. (Sequoias, some of which exceed 300 feet, are apparently near the upper limit of height for an organism.) Because of the buoyancy of water, large whales are not presented with such stability problems, but other difficulties arise. For a fixed shape, the volume of tissues to be nourished increases as the cube of the characteristic length of the organism, but the surface of the gut, which adsorbs the ingested food, increases only as the square of the length. As the length is increased, a point of diminishing returns is ultimately reached.

The range of organic molecules that organisms on Earth can metabolize is very wide and occasionally includes such foods as formaldehyde or petroleum, which seem unlikely from a human point of view. *Pseudomonas* bacteria are capable of using almost any organic molecule as a source of carbon and of energy, provided only that the molecule is at least slightly soluble in water. Microorganisms cannot metabolize plastics, not because of any fundamental chemical prohibitions but probably because plastics have not been part of the environment of microorganisms for very long. Man tends to think of oxygen as extremely important for life, but there are facultative anaerobes that can take their oxygen or leave it, and obligate anaerobes that are actually poisoned by oxygen. Such organisms use a variety of alternative electron acceptors, as previously discussed.

Chemical constituents

The water content of organisms usually represents between 50 and 90 percent of the live weight. Unless there is a massive mineral skeleton, the dry matter of organisms constitutes about one-half carbon by weight, reflecting the fact that organic molecules are based upon carbon. A wide variety of other chemical elements are used for diverse functions. Amino acids are made of nitrogen and sulfur in addition to carbon, hydrogen, and oxygen. Nucleic acids, as has been seen, employ phosphorus in addition to hydrogen, nitrogen, oxygen, and carbon. Sodium and potassium are used in maintaining the electrolyte balance, and calcium and silicon as structural materials. Iron plays a fundamental role in the transport of molecular oxygen as part of the hemoglobin molecule. In some ascidians (sea squirts), however, vanadium replaces iron. Ascidian blood also contains unusually large amounts of niobium, tantalum, titanium, chromium, manganese, molybdenum, and tungsten. The vanadium and niobium compounds in ascidian blood may be adaptations to low oxygen levels. Occasional organisms use selenium or tellurium as electron acceptors; others may produce the fully saturated gas hydrides of arsenic, phosphorus, or silicon, as metabolic wastes. Still others form compounds of carbon with such halogens as chlorine or iodine. Many of the foregoing elements, plus copper, zinc, cobalt, and possibly gallium, boron, and scandium, perform particular functions in the enzymatic apparatus of cells. Many of these elements, both the uncommon ones and those as common as phosphorus, are very highly concentrated in organisms over their general availability in the environment. This concentration must indicate that such chemicals play unique functional roles where other more abundant elements will not serve.

Behaviour and sensory capabilities. Analogous to the wide range of physiological adaptations and the great variety of elements used by organisms on Earth, there is an enormous range of behaviour patterns and sensory capabilities. Coded into its nucleic acids is the information that allows a bird raised from the egg in the absence of other birds to migrate when migration time arrives, to build a nest characteristic of its species, or to engage in elaborate courtship rituals. Those birds that do not perform acceptably do not leave descendants. Such behavioral information must itself have evolved. Rats that pass through mazes easily can be interbred, as can rats that pass through with difficulty; eventually two populations with inherited characteristics called "maze-smart" and "maze-dumb" will be produced. Fruit fly populations attracted to the light can be separated from those that avoid light. Classical genetic crossing experiments reveal that the two populations differ largely in a small number of genes for phototropism. Similar genetic determinants of behaviour exist in man. Possession of a supernumerary Y-chromosome in males is strikingly correlated with aggressive tendencies—which may, however, have been a selective advantage in more primitive societies. Myopia may have had strong survival value in earlier times: near-sighted males, useless in the hunt, stayed home and painted the walls of the cave. As technology develops, natural selection enters new behavioral arenas; for example, in an age of artificial contraception, the clumsy and forgetful preferentially reproduce.

Response to the electromagnetic spectrum

Human beings use only a small part of the total electromagnetic spectrum, the part called visible light, which extends from about 4,000 to about 7,000 Å in wavelength. While many plants and animals are sensitive to this same range of wavelengths, many of them are sensitive to other wavelengths as well. Most insects are sensitive to ultraviolet light at wavelengths below 4,000 Å, and many flowering plants take advantage of this fact and present patterns visible only in the ultraviolet range. Honeybees use polarized light, which the unaided human eye is quite unable to detect, for direction finding on partly cloudy days. The "pit" of such pit vipers as the rattlesnake is an infrared receptor and direction finder. These reptiles can sense the thermal radiation emitted by warm-blooded prey, radiation to which human beings are completely insensitive.

It is common knowledge that some animals (for example, dogs) are sensitive to sounds that the human ear cannot detect. Bats emit and detect sound waves at ultrahigh frequencies, in the vicinity of 100,000 cycles per second, about five times the highest frequency to which the human ear is sensitive. Bats use these sounds not so much to communicate, however, as to echolocate their prey and were doing this for millions of years before radar and sonar were invented. The audio receptors of many moths that are prey to bats are responsive only to the frequencies emitted by the bats. When the bat sounds are heard, the moths take evasive action. Dolphins have a very wide frequency range and several communication channels, as well as a "click" echolocator. Dolphins and whales use their blowholes rather than their mouths to utter these sounds. Sharks and other marine predators are said to locate their prey by the low-frequency sounds the prey makes when in distress. Some animals develop highly specialized and exotic organs for the detection or transmission of sound—for example, a European grasshopper has a relatively large parabolic antenna on its back that looks very much like a small radio telescope. This antenna is used for producing noises evidently thought attractive by the female of the species.

Many organisms are capable of smell and taste; that is, the detection of specific chemical molecules. According to one theory of smell, there are particular olfactory sensors, each receptive only to a specific chemical group on airborne molecules. The ultimate in olfactory specialization is probably the male silkworm moth: with its feathery antennae it is able to smell essentially nothing except the chemical sex attractant discharged by the female of the species. But it can detect this molecule very well, needing an impact of only 40 molecules per second on its antennae to produce a marked response. One female silkworm moth need release only 10^{-8} grams of sex attractant per second in order to attract every male silkworm moth in a volume hundreds of metres to kilometres on a side.

Extraordinary senses

Besides the senses of sight, hearing, smell, taste, and touch, various animals have a wide variety of other senses (see SENSORY RECEPTION). Man has an inertial orientation system and accelerometer in the cochlear canal of the ear. The water scorpion (*Nepa*) has a fathometer sensitive to hydrostatic pressure gradients. Most higher plants have chemically amplified gravity sensors. Fireflies and squids communicate with their own kind by producing time sequences or patterns of light on their bodies. The African freshwater fish *Gymnarchus niloticus* operates a dipole electrostatic field generator and a sensor to detect the amplitude and frequency of disturbances in the impressed field, an adaptation well suited for its nocturnal activities in turbulent waters. Other organisms have salinity sensors, or humidity sensors. There may be sensors involved in homing instincts of animals that have not yet been discovered. All of these senses confer upon their possessors an awareness of the environment that may be very different from that of such other organisms as man. Man, however, has the remarkable ability to extend his sensory and intellectual capabilities artificially, through the use of instrumentation.

THE ORIGIN OF LIFE

Hypotheses of origins. Perhaps the most fundamental and at the same time the least understood biological problem is the origin of life. It is central to many scientific and philosophical problems and to any consideration of extraterrestrial life. Most of the hypotheses of the origin of life will fall into one of four categories:

1. The origin of life is a result of a supernatural event; that is, one permanently beyond the descriptive powers of physics and chemistry.
2. Life—particularly simple forms—spontaneously and readily arises from nonliving matter in short periods of time, today as in the past.
3. Life is coeternal with matter and has no beginning; life arrived on the Earth at the time of the origin of the earth or shortly thereafter.
4. Life arose on the early Earth by a series of progressive chemical reactions. Such reactions may have been likely or may have required one or more highly improbable chemical events.

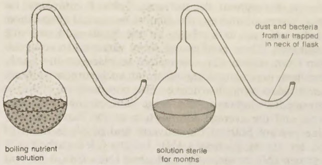
The early theological view

Hypothesis 1, the traditional contention of theology and some philosophy, is in its most general form not inconsistent with contemporary scientific knowledge, although this knowledge is inconsistent with a literal interpretation of the biblical accounts given in chapters 1 and 2 of Genesis and in other religious writings. Hypothesis 2 (not of course inconsistent with 1) was the prevailing opinion for centuries. A typical 17th-century view follows:

> [May one] doubt whether, in cheese and timber, worms are generated, or, if beetles and wasps, in cowdung, or if butterflies, locusts, shellfish, snails, eels, and such life be procreated of putrefied matter, which is to receive the form of that creature to which it is by formative power disposed[?] To question this is to question reason, sense, and experience. If he doubts this, let him go to Egypt, and there he will find the fields swarming with mice begot of the mud of the Nylus [Nile], to the great calamity of the inhabitants.

It was only in the Renaissance, with its burgeoning interest in anatomy, that such transformations were realized to be impossible. A British physiologist, William Harvey, during the mid-17th century, in the course of his studies on the reproduction and development of the king's deer, made the basic discovery that every animal comes from an egg. An Italian biologist, Francesco Redi, in the latter part of the 17th century, established that the maggots in meat came from flies' eggs, deposited on the meat. And an Italian priest, Lazzaro Spallanzani, in the 18th century, showed that spermatozoa were necessary for the reproduction of mammals. But the idea of spontaneous generation died hard. Even though it was proved that the larger animals always came from eggs, there was still hope for the smaller ones, the microorganisms. It seemed obvious that, because of their ubiquity, these microscopic creatures must be generated continually from inorganic matter.

Meat could be kept from going maggoty by covering it with a flyproof net, but grape juice could not be kept from fermenting by putting over it any netting whatever. This was the subject of a great controversy between the famous French bacteriologists Louis Pasteur and F.A. Pouchet in the 1850s, in which Pasteur triumphantly showed that even the minutest creatures came from germs floating in the air, but that they could be guarded against by suitable filtration. Actually, Pouchet was arguing that life must somehow arise from nonliving matter; if not, how had life come about in the first place?



Pasteur's swan-necked flask experiment (see text).

Life from outer space

Toward the end of the 19th century Hypothesis 3 gained currency, particularly with the suggestion by a Swedish chemist, S.A. Arrhenius, that life on Earth arose from panspermia, microorganisms or spores wafted through space by radiation pressure from planet to planet or solar system to solar system. Such an idea of course avoids rather than solves the problem of the origin of life. In addition, it is extremely unlikely that any microorganism could be transported by radiation pressure to the Earth over interstellar distances without being killed by the combined effects of cold, vacuum, and radiation.

Pasteur's work discouraged many scientists from discussing the origin of life at all. Moreover they were anxious not to offend religious feeling by probing too deeply into the subject. Although Darwin would not commit himself on the origin of life, others subscribed to Hypothesis 4 more resolutely, notably the famous British biologist T.H. Huxley in his *Protoplasm, the Physical Basis of Life* (1869), and the British physicist John Tyndall in his "Belfast Address" of 1874. Although Huxley and Tyndall asserted that life could be generated from inorganic chemicals, they had extremely vague ideas about how this might be accomplished. The very phrase "organic molecule" implies that there exists a special class of chemicals uniquely of biological origin, despite the fact that organic molecules have been routinely produced from inorganic chemicals since 1828. In the following discussion the word organic carries no imputation of biological origin. In fact the problem largely reduces to finding an abiological source of appropriate organic molecules.

The primitive atmosphere. Darwin's attitude was: "It is mere rubbish thinking at present of the origin of life; one might as well think of the origin of matter." The two problems are, in fact, curiously connected, and modern scientists are thinking about the origin of matter. There is convincing evidence that thermonuclear reactions and subsequent explosions in the interiors of stars generate all the chemical elements more massive than hydrogen and helium and then distribute them into the interstellar medium from which subsequent generations of stars and planets form. Because of the commonality of these thermonuclear processes, and because some thermonuclear reactions are more probable than others, there exists a cosmic distribution of the major elements, so far as is

Table 1: Relative Abundances of the Elements (percent)

atom	universe	life (terrestrial vegetation)	earth (crust)
Hydrogen	87	16	3
Helium	12	0*	0
Carbon	0.03	21	0.1
Nitrogen	0.008	3	0.0001
Oxygen	0.06	59	49
Neon	0.02	0	0
Sodium	0.0001	0.01	0.7
Magnesium	0.0003	0.04	8
Aluminum	0.0002	0.001	2
Silicon	0.003	0.1	14
Sulfur	0.002	0.02	0.7
Phosphorus	0.00003	0.03	0.07
Potassium	0.000007	0.1	0.1
Argon	0.0004	0	0
Calcium	0.0001	0.1	2
Iron	0.002	0.005	18

*0% here stands for any quantity less than 10^{-6}%.

known, throughout the universe. Table 1 compares, for some atoms of interest, the relative numerical abundances in the universe as a whole, on the Earth, and in living organisms. There is of course some variation in composition from star to star, from place to place on the Earth, and from organism to organism, but such comparisons are nevertheless very instructive. The composition of life is intermediate between the average composition of the universe and the average composition of the Earth. Ninety-nine percent both of the universe and of life is made of the six atoms, hydrogen (H), helium (He), carbon (C), nitrogen (N), oxygen (O), and neon (Ne). Can it be that life on Earth arose when the chemical composition of the Earth was much closer to the average cosmic composition, and that some subsequent events have changed the gross chemical composition of the Earth?

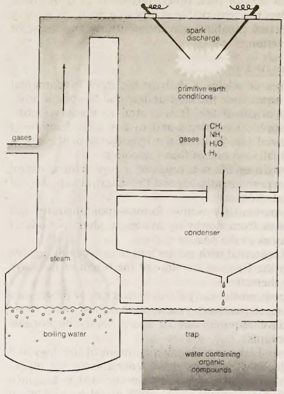
Composition of life

The Jovian planets (Jupiter, Saturn, Uranus, and Neptune) are much closer to cosmic composition than is the Earth. They are largely gaseous, with atmospheres composed principally of hydrogen and helium. Methane (CH_4) and ammonia (NH_3) have been detected in smaller quantities, and neon and water are suspected. This circumstance very strongly suggests that the Jovian planets were formed out of material of typical cosmic composition. They have very large masses, and because they are so far from the sun their upper atmospheres are very cold. Therefore it is impossible for atoms in the upper atmospheres of the Jovian planets to escape from their gravitational fields; escape was probably very difficult even during planetary formation. The Earth and the other planets of the inner solar system, however, are much less massive and most have hotter upper atmospheres. It is possible for hydrogen and helium to escape from the Earth today, and it may well have been possible for much heavier gases to have escaped during the formation of the Earth. It is reasonable to expect that in the very early history of the Earth a much larger abundance of hydrogen prevailed, which has subsequently been lost to space. Thus the atoms carbon, nitrogen, and oxygen were present on the primitive Earth, not as CO_2 (carbon dioxide), N_2, and O_2 as they are today but rather in the form of their fully saturated hydrides, CH_4 (methane), NH_3 (ammonia), and H_2O. In the geological record, the presence of such reduced minerals as uraninite (UO_2) and pyrite (FeS_2) in sediments formed several billions of years ago implies that conditions then were considerably less oxidizing than they are today.

In the 1920s J.B.S. Haldane in Britain and A.I. Oparin in the Soviet Union recognized that the abiological production of organic molecules in the present oxidizing atmosphere of the Earth is highly unlikely; but that, if the Earth once had more reducing (in this context, hydrogen-rich) conditions, the possible abiogenic production of organic molecules would have been much more likely. If large numbers of organic molecules were somehow synthesized on the primitive Earth, there would not necessarily be much trace of them today. In the present oxygen atmosphere, largely produced by green-plant photosynthesis, such molecules would tend, over geological time, to be oxidized to carbon dioxide, nitrogen, and water. In addition, as Darwin recognized, the first microorganisms would consume prebiological organic matter produced prior to the origin of life.

Synthesis of amino acids

Production of simple organic molecules. The first deliberate experimental simulation of these primitive conditions was carried out in 1953 by a U.S. graduate student, S.L. Miller, under the guidance of the eminent chemist H.C. Urey. A mixture of methane, ammonia, water vapour, and hydrogen was circulated through a liquid water solution and continuously sparked by a corona discharge elsewhere in the apparatus. The discharge may be thought to represent lightning flashes on the early Earth. After several days of exposure to sparking, the solution changed colour. Subsequent analysis indicated that several amino and hydroxy acids, intimately involved in contemporary life, had been produced by this simple procedure. The experiment is in fact so elementary, and the amino acids can so readily be detected by paper chromatography, that the experiment has been repeated many times by high school students. Subsequent experiments have substituted ultraviolet light or heat as the energy source or have altered the initial abundances of gases. In all such experiments amino acids have been formed in large yield. On the early Earth there was much more energy available in ultraviolet light than in lightning discharges. At long ultraviolet wavelengths, in which methane, ammonia, water, and hydrogen are all transparent, but in which the bulk of the solar ultraviolet energy lies, the gas hydrogen sulfide (H_2S) is a likely ultraviolet absorber.



Miller–Urey spark-discharge apparatus.

Following such reasoning, a U.S. astrophysicist, Carl Sagan, and his colleagues made amino acids by long wavelength ultraviolet irradiation of a mixture of methane, ammonia, water, and H_2S. The amino acid syntheses, at least in many cases, involve hydrogen cyanide and aldehydes (*e.g.*, formaldehyde) as gaseous intermediaries formed from the initial gases. It is quite remarkable that amino acids, particularly biologically abundant amino acids, can be made so readily under simulated primitive conditions. When laboratory conditions become oxidizing, however, no amino acids are formed, suggesting that reducing conditions were necessary for prebiological organic synthesis.

Under alkaline conditions, and in the presence of inorganic catalysts, formaldehyde spontaneously reacts to form a variety of sugars, including the five-carbon sugars fundamental to the formation of nucleic acids and such six-carbon sugars as glucose and fructose, which are extremely common metabolites and structural building blocks in contemporary organisms. Furthermore, the nucleotide bases as well as porphyrins have been produced in the laboratory under simulated primitive Earth conditions by several investigators. While there is still debate on the generality of the experimental synthetic pathways and on the stability of the molecules produced, most if not all of the essential building blocks of proteins, carbohydrates, and nucleic acids can be readily produced under quite general primitive reducing conditions, plus probably ATP as well.

Polymerization as dehydration reactions

Production of polymers. The construction of polymers, long-chain molecules made of repeating units of these essential building blocks, however, is a much more difficult experimental problem. Polymerization reactions are generally dehydrations, in which a molecule of water is lost in the formation of a two-unit polymer. Dehydrating agents must be used to initiate polymerization. The polymerization of amino acids to form long protein-like molecules was accomplished through dry heating by a U.S. investigator, S.W. Fox. The polyamino acids that are formed are not random polymers and have some distinct catalytic activities. The geophysical generality of dry heating and return to solution, however, has been questioned. Long polymers of amino acids can also be produced from hydrogen cyanide and anhydrous liquid ammonia. Some evidence exists that nucleotide bases and sugars can be combined in the presence of phosphates or cyanides under ultraviolet irradiation. Some condensing agents such as cyanamide are efficiently made under simulated primitive conditions. Despite the breakdown by water of molecular intermediates, condensing agents are often quite effective in inducing polymerization, and polymers of amino acids, sugars, and nucleotides have all been made this way.

A famous British scientist, J.D. Bernal, suggested that adsorption of molecular intermediates on clays or other minerals may have concentrated these intermediates. Such concentration could offset the tendency for water to break down polymers of biological significance. Of special interest is the possibility that such concentration matrices included phosphates, for this would help explain how phosphorus could have been incorporated preferentially into prebiological organic molecules at a time when biological concentration mechanisms did not yet exist. Mineral catalysis implies that organic synthesis could also occur in deep water where ultraviolet light had been filtered out.

Quite apart from concentration mechanisms, the primitive waters themselves may have been a not very dilute solution of organic molecules. If all the surface carbon on the Earth were present as organic molecules in the contemporary oceans, or if many known ultraviolet synthetic reactions producing organic molecules were permitted to continue for a billion years with products dissolved in the oceans, a 1 percent solution of organic molecules would result. For similar reasons, Haldane suggested that the origin of life occurred in a "hot dilute soup." In addition, concentration mechanisms do exist, such as evaporation or freezing of pools or adsorption on interfaces or the generation of colloidal enclosures called coacervates.

The origin of the code. It has been shown that all the essential building blocks for life and their polymers may have been produced in some fair concentration on the primitive Earth. This possibility is certainly relevant to the origin of life, but it is not the same thing as the origin of life. By the genetic definition of life discussed above in *Definitions of life,* a self-replicating, mutable molecular system, capable of interacting with the environment, is required. In contemporary cells the nucleic acids are the sites of self-replication and mutation. Laboratory experiments have already shown that polynucleotides can be produced from nucleotide phosphates in the presence of a specific enzyme of biological origin and a pre-existing "primer" nucleic acid molecule. If the primer molecule is absent, polynucleotides are still formed, but they of course contain no genetic information. Once such a polynucleotide spontaneously forms, it then acts as primer for subsequent syntheses.

Production of polynucleotides

Imagine a primitive ocean filled with nucleotides and their phosphates and appropriate mineral surfaces serving as catalysts. Even in the absence of the appropriate enzyme it seems likely, although not yet proved, that spontaneous assembly of nucleotide phosphates into polynucleotides occurred. Once the first such polynucleotide was produced, it may have served as a template for its own reproduction, still of course in the absence of enzymes. As time went on there were bound to be errors in replication. These would be inherited. A self-replicating and mutable molecular system of polynucleotides, eventually leading to a diverse population of such molecules, may have arisen in this way. Alternatively, the primitive hereditary material may have involved some other molecule altogether, but no concrete suggestion for such a molecule has ever been proposed.

In any case, a population of replicating polynucleotides cannot quite be considered alive because it does not significantly influence its environment. Eventually, all the nucleotides in the ocean would have been tied in polynucleotides and the entire synthetic process would then have ground to a halt. So far as is known, polynucleotides have no an catalytic properties, and proteins have no reproductive properties. It is only the partnership of the two molecules that makes contemporary life on Earth possible. Accordingly, a critical and unsolved problem in the origin of life is the first functional relation between these two molecules, or, equivalently, the origin of the genetic code. The molecular apparatus ancillary to the operation of the code—the activating enzymes, adapter RNAS, messenger RNAS, ribosomes, and so on—are themselves each the product of a long evolutionary history and are produced according to instructions contained within the code. At the time of the origin of the code such an elaborate molecular apparatus was of course absent.

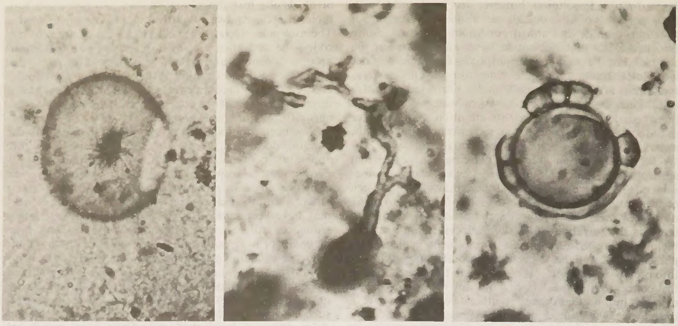
Catalysis and reproduction as necessary for life

It has been proposed that a weak but selective chemical bonding does exist, even in the absence of any of this apparatus, between amino acids and nucleotides. There need not be a very great selectivity; a given nucleotide sequence might in primitive times have coded for many different amino acids or, conversely, the same amino acid may have been coded for by several different nucleotide sequences. All that is required is that a particular linear sequence of nucleotides must code for some nonrandom sequence of amino acids. The active sites largely responsible for the catalytic activity of contemporary enzymes are generally only five or six amino acids long; the remainder of the enzyme is devoted to more sophisticated functions, such as arranging for the enzyme to be turned on and off by the machinery of the cell. With, say, 20 different varieties of amino acids available in the primitive environment, the chance of any given active site being produced by a random sequence of nucleotides is one in 20^5, or one in about 3,000,000. But 3,000,000 combinations to form units five amino acids long is not a very large number for the chemistry and time periods in question. To conclude this speculation, then, if polynucleotides were initially capable of crude, nonenzymatic replication, and if a crude primitive genetic code existed, then any one of a very large number of catalytic properties was available to some self-replicating polynucleotides on the primitive Earth. This situation is all that would be necessary for the origin of life; those polynucleotides that could code for a primitive protein having catalytic properties furthering the replication of the polynucleotide would preferentially replicate. Other polynucleotides coding for less effective proteins would have replicated more slowly. The foregoing is one of several possibilities for the origin of the first living systems. Many separate and rather diverse instances of the origin of life may have occurred on the primitive Earth, but competition eventually eliminated all but one line. Every organism on Earth today would be a descendant of that line.

The earliest living systems. One curious feature of biological organic molecules is their optical activity: they rotate the plane of a beam of plane-polarized light. Organic molecules produced abiologically do not show optical activity. Molecules made of the same units can be put together in complementary ways like a left- and right-handed glove. The same building blocks can be used to produce molecules that are three-dimensional mirror images of each other. This asymmetry is responsible for optical activity. At the time of the origin of life, organic molecules, corresponding both to left- and right-handed forms, were produced. The laboratory simulation experiments always produce both types. But the first living systems could have been made only of one type, for the same reason that carpenters do not use random mixtures of screws with left- and right-handed threads. Whether left- or right-handed activity was adopted was probably purely a matter of chance, but once a particular asymmetry was established in the first living systems, it maintained itself. This belief implies that optical activity should be a feature of life on any planet, and also that the chances should be equal of finding a given terrestrial organic molecule or its mirror image molecule in extraterrestrial life forms.

Optical activity of biological molecules

The first living systems probably resided in a molecular garden of Eden, where all the building blocks that contemporary organisms must work hard at synthesizing were available free. Under such conditions the numbers of organisms must have increased very rapidly. But such increases cannot go on indefinitely. In time the supply of some molecular building block must have become short. Those primitive organisms that had the ability to synthesize the scarce building block, say A, from a more



Photomicrographs of 2,000,000,000-year-old organisms from the geological stratum called the Gunflint Chert. These organisms are about 10 microns across. Organisms more than 1,000,000,000 years older are known.
By courtesy of Elso S. Barghoorn

abundant one, say B, clearly had a competitive advantage over those organisms that could not perform such a synthesis. In time, however, the secondary source of supply, B, would have also become depleted and those organisms that could produce it from a third building block, C, would have preferentially replicated. A U.S. biochemist, N.H. Horowitz, proposed that in this way the enzymatic reaction chains of contemporary organisms—each step catalyzed by a particular enzyme—originally evolved.

Even the evolution of enzymatic reaction chains may have occurred in free nucleic acids before the origin of the cell. The cell may have arisen in response to the need for maintaining a high concentration of scarce building blocks or enzymes, or as protection against the gradually increasing abundance of oxygen on the primitive Earth. Oxygen is a well-known poison to many biological processes, and in contemporary higher organisms the mitochondria that handle molecular oxygen are kept in the cytoplasm, far from contact with the nuclear material. Even today processes are known whereby polyamino acids form small spherical objects, microns to tens of microns across, with some of the properties of cells. These objects, called proteinoid microspheres by Fox, are certainly not cells, but they may indicate processes by which the ancestors of cells arose. Procaryotic cells almost certainly preceded eucaryotic cells, and the evolution of so extremely complex an apparatus as the mitotic spindle (which ensures equal segregation of replicated chromosomes) must have taken very long periods of time to evolve. The development of mitochondria and chloroplasts (each of which contains its own DNA) in the eucaryotic cell may have been the result of a symbiosis, a cooperative arrangement entered into at first tentatively by originally free-living cells.

Proteinoid microspheres

As the competition for building blocks increased among early life forms, and also perhaps as the abiological production of organic molecules dwindled because of the increasing oxygen abundance, the strictly heterotrophic way of life became more and more costly. The utilization of porphyrins, which are also made abiologically, by primitive photoautotrophs would have had great selective advantage. Many of the intermediates and enzymes in photosynthesis and in the anaerobic breakdown of carbon compounds are similar, but there is no generally accepted view of the origin of the photosynthetic process. Photosynthesis in procaryotes is more primitive than in such eucaryotes as green plants. In bacteria, water is not the ultimate source of hydrogen atoms for reducing carbon dioxide, and therefore oxygen is not produced. In addition, when a chlorophyll-containing cell is exposed both to light and to oxygen, it is killed unless it also contains an accessory carotenoid pigment. Thus green-plant photosynthesis had to wait until the appearance of carotenoids while bacterial photosynthesis, which does not produce oxygen, could function without carotenoids.

The antiquity of life. Among the oldest known fossils are those found in the Fig Tree chert from the Transvaal, dated at 3,100,000,000 years old. These organisms have been identified as bacteria and blue-green algae. It is very reasonable that the oldest fossils should be procaryotes rather than eucaryotes. Even procaryotes, however, are exceedingly complicated organisms and very highly evolved. Since the Earth is about 4,500,000,000 years old, this suggests that the origin of life must have occurred within a few hundred million years of that time.

By performing chemical analyses on the oldest sediments, it is possible to say something about the sorts of organic molecules produced, either biologically or abiologically, in primitive times. Thus, amino acids and porphyrins have been identified in the oldest sediments, as have pristane and phytane, typical breakdown products of chlorophyll. There are several indications that these organic molecules, dating from 2,000,000,000 to more than 3,000,000,000 years ago, are of biological origin. For one thing their long-chain hydrocarbons show a preference for a straight chain geometry, whereas known abiological processes tend to produce a much larger proportion of branched chain and cyclic hydrocarbon molecular geometries than have been found in these sediments. Abiological processes tend to produce equal amounts of long-chain carbon compounds with odd and with even numbers of carbon atoms. But the oldest sediments show a distinct preference for odd numbers of carbon atoms per molecule, as do products of undoubted biological origin. Finally, a C^{12} enrichment, for which no abiological process seems able to account, has been discovered in the oldest sediments, evidence that suggests that plantlike life, which concentrates the carbon isotope C^{12} preferentially to C^{13}, was present very early. These departures from thermodynamic equilibrium are often considered to be compelling signs of biological activity. Such evidence again points to the great antiquity of life on Earth.

The fossil record, in any complete sense, goes back only about 600,000,000 years. In the layers of sedimentary rock known by geological methods and by radioactive dating to be that old, most of the major groups of invertebrates appear for the first time. All these organisms appear adapted to life in the water, and there is no sign yet of organisms adapted to the land. For this reason, and because of a rough similarity between the salt contents of blood and of seawater, it is believed that early forms of life developed in oceans or pools. With no evidence for widespread oxygen-producing photosynthesis before this time, and for cosmic

Oceanic origin of life

abundance reasons described above, the oxygen content of the Earth's atmosphere in Precambrian times was very likely less than today. Accordingly, in Precambrian times, solar ultraviolet radiation, especially near the wavelength of 2,600 Å, which is particularly destructive to nucleic acids, may have penetrated to the surface of the Earth, rather than being totally absorbed in the upper atmosphere by ozone as it is today. In the absence of ozone, the ultraviolet solar flux is so high that a lethal dose for most organisms would be delivered in less than an hour. Unless extraordinary defense mechanisms existed in Precambrian times, life near the Earth's surface would have been impossible. Sagan suggested that life at this time was generally restricted to some tens of metres and deeper in the oceans, at which depths all the ultraviolet light would have been absorbed, although visible light would still filter through. As the amount of atmospheric oxygen and ozone increased, due both to plant photosynthesis and to the photodissociation of water vapour and the escape to space of hydrogen from the upper atmosphere, life increasingly close to the Earth's surface would have been possible. It has been suggested that the colonization of the land, about 425,000,000 years ago, was possible only because enough ozone was then produced to shield the surface from ultraviolet light for the first time.

Life then had insinuated itself between the sun and the Earth. It diverted solar energy to its own uses and contrived more and more ways of exploiting more and more environments. Some experiments were faulty and the lines became extinct; others were more successful and the lines filled the Earth. Evolution through natural selection directed the proliferation of a growing array of life forms throughout the biosphere (see EVOLUTION, THE THEORY OF).

EXTRATERRESTRIAL LIFE

It is not known what aspects of living systems are necessary in the sense that living systems everywhere must have them; it is not known what aspects of living systems are contingent in the sense that they are the result of evolutionary accident, so that somewhere else a different sequence of events might have led to different characteristics. In this respect the possession of even a single example of extraterrestrial life, no matter how seemingly elementary in form or substance, would represent a fundamental revolution in biology. It is not known whether there is a vast array of biological themes and counterpoints in the universe, whether there are places that have fugues, compared with which our one tune is a bit thin and reedy. Or it may be that our tune is the only tune around. Accordingly the prospects for life on other planets must be considered in any general discussion of life.

The chemistry of extraterrestrial life. What are the methods and prospects for a search for life beyond the Earth? Each of the definitions of life described in *Definitions of life* (see above) implies a method of searching for life. Particular physiological functions, particular metabolic activities, such specific molecules as proteins and nucleic acids, self-replication and mutation, processes not in closed-system thermodynamic equilibrium—all these might be sought. All the search methods significantly depend upon chemistry.

Life on Earth is structurally based on carbon and utilizes water as an interaction medium. Hydrogen and nitrogen have significant accessory structural roles; phosphorus is important for energy storage and transport, sulfur for three-dimensional configuration of protein molecules, and so on. But must these particular atoms be the atoms of life everywhere, or might there be a wide range of atomic possibilities in extraterrestrial organisms? What are the general physical constraints on extraterrestrial life?

Prerequisites for life

In approaching these questions several criteria can be used. The major atoms should tend to have a high cosmic abundance. A structural molecule for making an organism at the temperature of the planet in question should not be extremely stable, because then no chemical reactions would be possible; but it should not be extremely unstable, because then the organism would fall to pieces. There should be some medium for molecular interaction. Solids are not appropriate because the diffusion times are very long. Such a medium is most likely a liquid (but could possibly be a very dense gas) that is stable in a number of respects. It should have a large temperature range (for a liquid, the temperature difference between freezing point and boiling point should be large). The liquid should be difficult to vaporize and to freeze; in fact, it should be very difficult to change its temperature at all. In addition it should be an excellent solvent. There should also be some gas on the planet in question that could be used in various biologically mediated cycles, as CO_2 is in the carbon cycle on Earth.

The planet, therefore, should have an atmosphere and some near-surface liquid, although not necessarily an ocean. If the intensity of ultraviolet light or charged particles from the sun is intense at the planetary surface, there must be some place, perhaps below the surface, that is shielded from this radiation but that nevertheless permits useful chemical reactions to occur. Since after a certain period of evolution, lives of unabashed heterotrophy lead to malnutrition and death, autotrophs must exist. Chemoautotrophs are, of course, a possibility but the inorganic reactions that they drive usually require a great deal of energy; at some stage in the cycle, this energy must probably be provided by sunlight. Photoautotrophs, therefore, seem required. Organisms that live very far subsurface will be in the dark, making photoautotrophy impossible. Organisms that live slightly subsurface, however, may avoid ultraviolet and charged particle radiation and at the same time acquire sufficient amounts of visible light for photosynthesis.

Thermodynamically, photosynthesis is possible because the plant and the radiation it receives are not in thermodynamic equilibrium; for example, on the Earth a green plant may have a temperature of about 300 K while the sun has a temperature of about 6,000 K. (K = Kelvin temperature scale, in which 0 K is absolute zero; 273 K, the freezing point of water; and 373 K, the boiling point of water at one atmosphere pressure.) Photosynthetic processes are possible in this case because energy is transported from a hotter to a cooler object. Were the source of radiation at the same (or at a colder) temperature as the plant, however, photosynthesis would be impossible. For this reason the idea of a subterranean plant photosynthesizing with the thermal infrared radiation emitted by its surroundings is untenable, as is the idea that a cold star, with a surface temperature similar to that of the Earth, would harbour photosynthetic organisms.

Chemical requirements

It is possible to approach some of the foregoing chemical requirements and see just which atoms are implied. When atoms enter into chemical combination, the energy necessary to separate them is called the bond energy, a measure of how tightly the two atoms are bound to each other. Table 2 gives the bond energies of a number of chemical bonds, mostly involving abundant atoms. The energies are in electron volts (eV; 1 eV = 1.6×10^{-12} ergs). The symbols are as follows: H, hydrogen; C, carbon; N, nitrogen; O, oxygen; S, sulfur; F, fluorine; Si, silicon; Bi, bismuth (very underabundant, biologically uninteresting, and present only as an illustration of the relatively weak chemical bonds in some metals). Bond energies generally vary between 10 eV and about 0.03 eV; double and triple

Table 2: Energies of Representative Chemical Bonds

(Hydrogen bonds 0.08–0.45 eV; van der Waals bonds 0.04 eV)

bond	energy (eV)	bond	energy (eV)
N≡N	9.8	Si—O	3.8
C≡N	9.4	C—O	3.7
C≡C	8.4	C—C	3.6
C—O	7.4	S—H	3.5
C=C	6.4	Si—H	3.1
H—F	5.4	C—N	3.0
O—H	4.8	Si—Si	1.8
N=N	4.4	N—N	1.7
C—H	4.3	Bi—Bi	1.1
N—H	4.1	$O_2N—NO_2$	0.57

bonds where two or three electrons are shared between two atoms tend to be more energetic than single bonds, single bonds more energetic than hydrogen bonds where a hydrogen atom is shared between two other atoms, and hydrogen bonds more energetic than the very weak (van der Waals) forces that arise from the attraction of the electrons of one atom for the nucleus of another. At room temperature, atoms, free or bound, move with an average kinetic energy corresponding to about 0.02 eV. Some of the atoms have greater energies, some lesser. At any temperature a few will have energies greater than any given bond energy; hence bonds occasionally will break. The higher the temperature, the more atoms there are moving with sufficient energy to spontaneously break a given bond.

Suppose it is decided arbitrarily (although the decision will not critically affect the conclusions) that for life to exist at any time the fraction of bonds broken by random thermal motions must be no larger than 0.0001 percent. It then turns out that any hypothetical life where the structural bonds are based upon van der Waals forces can only exist where the temperature is below 40 K, for hydrogen bonds below about 400 K, for bonds of 2 eV below 2,000 K, and for bonds of 5 eV below 5,000 K. Now, 2,000 to 5,000 K are typical surface temperatures of stars; 400 K is somewhat above the highest surface temperature found on Earth; and 40 K is about the cloud-top temperature of distant Neptune. Thus, over the entire range of temperatures, from cold stars to cold planets, there seem to exist chemical bonds of appropriate structural stability for life, and it would appear premature to exclude the possibility of life on any planet on grounds of temperature.

Life on Earth lies within a rather narrow range of temperature. Above the normal boiling point of water, much loss of configurational structure or three-dimensional geometry occurs. At these temperatures proteins become denatured, in part because above the boiling point of water the hydrogen bonding and van der Waals forces between water and the protein disappear. Also, similar bonds within the protein molecule tend to break down. Proteins then change their shapes, their ability to participate in lock-and-key enzymatic reactions is gravely compromised, and the organism dies. Similar structural changes, some of them connected with the stacking forces between adjacent nucleotide bases, occur in the heating of nucleic acids. But it is significant that these changes are not fragmentations of the relevant molecules but rather changes in the ways they fold. There appears to be no reason that configurational bonds should not have been evolved that are stable at higher temperatures than terrestrial organisms experience. On planets hotter than the Earth there seems to be no reason that slightly more stable configurational forces should not be operative in the local biochemistry.

Molecular factors

While the bonds that characterize life on Earth are too weak at high temperatures, they are too strong at low temperatures, tending to slow down the rates of chemical reactions generally. There are less stable bonds (*e.g.*, hydrogen bonds, silicon-silicon bonds, and nitrogen-nitrogen bonds), however, that might play structural roles at significantly lower temperatures. At higher temperatures, multiple bonds (*e.g.*, in aromatic, or ring-shaped, hydrocarbons) might be utilized for life. There clearly is a rich variety of little-studied chemical reactions that proceed at reasonable rates either at much lower or at much higher temperatures than those on Earth.

Except for bismuth and fluorine, all the atoms in Table 2 have relatively high cosmic abundances. At terrestrial temperatures, carbon is the unique atom for biological structure. Not only does it have high abundance but it forms a staggering variety of compounds of great stability, it lends itself to compounds that are configured by weaker bonds, and it enters into multiple bonds. These double- and triple-bonded molecules, among other useful properties, absorb long-wavelength ultraviolet light, a process leading to the synthesis of a variety of more complex molecules. A photon of ultraviolet light at a wavelength of 2,000 Å has an energy of 6.2 eV, capable of breaking many bonds, and permitting more complex reactions among the resulting molecular fragments. Photons of blue light have energies of about 3 eV, and of red light about 2 eV.

Silicon compounds do not form double bonds. Silicon-oxygen bonds are slightly more stable than carbon-carbon bonds, but they tend to produce molecules like the silicates, which are crystals of the same unit repeated over and over again, rather than molecules with aperiodic side chains with potential information content. On low-temperature planets, silicon-silicon bonds are more promising than carbon bonds in terms of reaction times, but they do not form double bonds and the carbon abundance is likely to be greater. Nevertheless, silicon compounds may be of limited biological importance both on high-temperature and low-temperature worlds.

Hydrogen bonding confers on liquids the stability properties necessary for life. There seem to be very few reasonable candidates for liquid interaction media. By all odds water is the most suitable. The other candidates, all to some extent hydrogen bonded, are ammonia, hydrogen fluoride, hydrogen cyanide, and mixtures of liquid hydrocarbons. Hydrogen fluoride can be excluded because it is too scarce cosmically. The hydrocarbons are not good solvents of salts, but life elsewhere may not be based on the same acid-base chemistry as life on Earth. The liquid range of water is larger than commonly thought, ranging from about 210 K in saturated salt solutions to 647 K at enormous atmospheric pressures. Water is the biological liquid medium of choice above 200 K, particularly in view of its extremely high cosmic abundance. At lower temperatures ammonia or hydrogen cyanide could serve as a liquid medium.

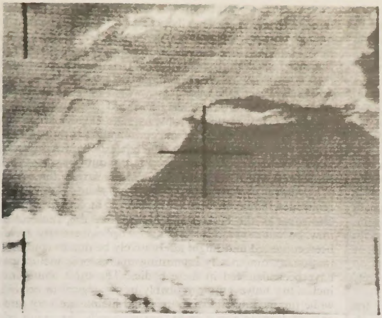
There are functional roles for specific atoms in biology, but except for considerations of structure and a liquid interaction medium they do not seem fundamental. For example, the energy-rich phosphate bonds in ATP are in fact of relatively low energy; they are about as energetic as the hydrogen bonds (see Table 2). The cell must store up large numbers of these bonds to drive a molecular degradation or synthesis. On high-temperature worlds the energy currency may be much more energetic per bond, and on low-temperature worlds much less energetic per bond.

It may be concluded that, in our present state of ignorance, it is premature to exclude life on grounds of temperature on any other planet, particularly when account is taken of the temperature heterogeneity of the other planets. But life does require an interaction medium, an atmosphere, and some protection from ultraviolet light and from charged particles of solar origin.

The conclusion that for the Earth, carbon-based aqueous life is the most appropriate may be slightly suspect, since terrestrial life is manifestly carbon-based and aqueous. In 1913 a U.S. biochemist, L.J. Henderson, published *The Fitness of the Environment* in which the biological advantages of carbon and water were stressed for the first time in terms of comparative chemistry. He was struck by the fact that those very atoms that are needed are just those atoms that are around; it remains a remarkable fact that atoms most useful for life do have very high cosmic abundances.

Exobiology

The search for extraterrestrial life. Exobiology, a term coined by a U.S. biologist, J. Lederberg, for the study of extraterrestrial life, has been called a science without a subject matter. It is certainly true that, as yet, no strong evidence for life beyond the Earth has been adduced. Exobiology, however, has deep significance even if extraterrestrial life is never found. The mere design of exobiological experiments forces man to examine critically the generality of his assumptions about life on Earth. In addition, a lifeless neighbouring planet presents a very interesting quandary: How is it that life has originated and evolved on Earth, but not on the planet in question? There is an entire spectrum of possibilities. A given planet may be lifeless and have no vestiges of primitive organic matter and no fossils of extinct life. It may be lifeless but may have either organic chemical or fossil relics. It may possess life of a simple sort or life of a quite complex biochemistry, physiology, and behaviour. It may possess intelligent life and a technical civilization. Establishment of any one of these five possibilities would be of fundamental biological importance.



(Left) Eastern seaboard of the United States photographed by a TIROS weather satellite; Cape Cod (right), Long Island (centre), and Delaware Bay (centre left) can be seen. (Right) Surface of Mars photographed by Mariner 4 spacecraft. Both photographs at one-kilometre resolution. No sign of life, intelligent or otherwise, can be discerned on either planet.

By courtesy of National Aeronautics and Space Administration

The difficulties and opportunities inherent in exobiological exploration, in determining which of these five possibilities applies to a given planet, is most clearly grasped by imagining the situation reversed, with man on some neighbouring planet, say Mars, examining the Earth for life with the full armoury of contemporary scientific instrumentation and knowledge. First a distinction must be made between remote and *in situ* testing. In remote testing light of any wavelength reflected from or emitted by the target planet can be examined, but with *in situ* studies samples of the planet must be acquired by visiting them or by sending instruments that land on the planet, perform experiments, and radio back their findings. Since biological exploration involves the detailed characterization of any life found, rather than its mere detection, *in situ* experiments are necessary.

Sensing methods

The bulk of the remote sensing methods are directed toward finding some thermodynamic disequilibrium on the planet. This may be a chemical disequilibrium, a mechanical disequilibrium, or a spectral disequilibrium. For example, it would be quite easy to determine spectroscopically from Mars that the Earth's atmosphere contains large amounts of molecular oxygen and about one part per million (10^6) of methane. It would also be possible to calculate that, at thermodynamic equilibrium, the abundance of methane should be less than one part in 10^{35}. This huge discrepancy implies the existence of some process continuously generating methane on the Earth so rapidly that methane increases to a very large steady-state abundance before it can be oxidized by oxygen. Now such a methane-production mechanism need not be biological. It is conceivable that relatively stable aromatic hydrocarbons were produced abiologically in the early history of the Earth and that their slow thermal degradation leads to a continuous loss of methane from the planetary subsurface. But this and similar nonbiological explanations of the observed disequilibrium are unlikely. From Mars this thermodynamic discrepancy would be considered not as proof of life on Earth but as a significant hint of life on Earth. In fact the methane abundance on the Earth is produced by bacteria that, in the course of the reduction of a more oxidized form of carbon, release methane. Some methane bacteria live in swamps (hence, the term marsh gas for methane), and others—a significant fraction—live in the intestinal tracts of cows and other ruminants. The methane abundance over India is probably larger than over most other areas of the world, and if an extraterrestrial observer knew how to interpret the methane disequilibrium accurately (which is unlikely) it would be possible for him to deduce cows on Earth by spectrochemical analysis. The existence of relatively large quantities of methane in the presence of an excess of oxygen would remain a tantalizing but enigmatic hint of life on Earth. Similarly, the large amount of oxygen might itself be a sign of life if one could reliably exclude the possibility that the photodissociation of water and the escape to space of hydrogen were the source of oxygen. Also such relatively complex reduced organic molecules as terpenes, a hydrocarbon given off by plants, might conceivably be detected spectroscopically, perhaps by a spectrometer in orbit about the Earth. Not only would the chemical disequilibrium of terpenes in an excess of oxygen be suggestive of life, but equally suggestive would be the fact that terpenes are much more abundant over forested areas than over deserts.

Photographic observation

Photographic observations of the daytime Earth from Mars would give equivocal results. Even with a resolution of 100 metres (that is, an ability to discriminate fine detail at high contrast only if its components are more than 100 metres apart), it would be extremely difficult to discern cities, canals, bridges, the Great Wall of China, highways, and other large-scale accoutrements of the Earth's technical civilization. In satellite photographs with 100-metres (one metre = 1.0936 yards) resolution only about one in a thousand random photographs of the Earth yields features even suggestive of life. As the ground resolution is progressively improved, it becomes increasingly easy to make out the regular geometrical patterns of cultivated fields, highways, airports, and so on. But these are only the products of a civilization recently developed on Earth, and even photographs of the Earth with a ground resolution of 10 metres, but taken 100,000 years ago, would still have shown no clear sign of life. The lights of the largest cities might be just marginally detectable from Mars at night. Seasonal changes in the colour or darkness of plants would be detectable from Mars, but such cycles might easily have nonbiological explanations.

To detect individual animals a ground resolution of a few metres is required, and even here a low sun and long shadows are generally necessary. This detection could be accomplished with a large telescope in Earth orbit. It would then be possible to determine, for example, that objects with the general shape of cows are frequent on the Earth. But suppose that members of the civilization examining the Earth thus remotely are not even approximately quadrupedal and do not immediately associate the shape of cows with life. They would nevertheless be able to deduce life. They would observe that certain locales on Earth have a quantity of raised lumps connected to the ground by four stilts. It would be possible to calculate that wind and water erosion would cause the lumps to topple to the ground in geologically short periods of time. Such stilted lumps are mechanically unstable; they are not in equilibrium; if pushed hard, they fall. Accordingly, there must be a process for generating stilted lumps on the Earth in short periods of time. It would be very difficult to avoid the implication that this generation process is biological.

Radio emission

A third detection technique arises upon scanning the radio spectrum of the Earth. Because of domestic television transmission, the high-frequency end of the AM broadcast band, and the radar defense networks of the United States and various other countries, the amount or energy put out by the Earth to space at certain radio frequencies is enormous. At some frequencies, if this radiation were to be interpreted as ordinary thermal emission, the temperature of the Earth would have to be hundreds of millions of degrees, according to an estimate made by a Russian astrophysicist, I.S. Shklovskii. Moreover, it would be possible to determine that this radio "brightness temperature" of the Earth had been steadily increasing with time over the last several decades. Finally, it would be possible to analyze the frequency and the time variation of these signals and deduce that they were not purely random noise.

Now imagine in situ studies by vehicles that enter the Earth's atmosphere and land at some predetermined locale. There are many places on the Earth (the ocean surface, the Gobi Desert, Antarctica) where large organisms are infrequent and a life-detection attempt based solely on television searches for large life forms would be a risky investment. On the other hand, if such an experiment were successful (the camera records a dolphin cavorting, a camel chewing its cud, a penguin waddling), it would provide quite convincing evidence of life.

Although the oceans, the Gobi Desert, and Antarctica are relatively devoid of large life forms, they are in many places replete with minute life forms. Therefore, microorganism detectors would be a good investment. A television camera coupled to a microscope (optical or electron) would be a promising life detector if the sample acquisition problem could be solved: the early Dutch microscopist Antonie van Leeuwenhoek had no difficulty at all in identifying as alive the little "animalcules" that he found in a drop of water, although nothing similar had previously been seen in human history.

In addition to morphological criteria for the detection of microorganisms, there are metabolic and chemical criteria. For example, a sample of terrestrial soil, or seawater, say, might be acquired and introduced into a chamber containing food the investigators guess the earthlings might find tasty. Such food might be an abundant product of prebiological organic synthetic experiments. It could then be determined whether any characteristic molecules, such as carbon dioxide or ethanol, are produced metabolically or whether the medium containing food and terrestrial sample changes its acidity or becomes cloudy because of the growth of microorganisms, or it might be investigated whether there is heat given off in the chamber containing sample and food. Alternatively, photosynthesis could be tested by measuring the fixation of some gas, say carbon dioxide, as a function of illumination provided artificially to the sample by the instrument. Along chemical lines a direct test of terrestrial soil or seawater for optical activity might be made. Organic molecules could certainly be searched for with a combined gas chromatograph and mass spectrometer or by a remote analytic chemistry laboratory. The detection of any amount of organic matter would of course be interesting and relevant, whether or not it was biological in origin. Such criteria as have been used in the analysis of Precambrian sediments (described in *The antiquity of life,* above) might be used to test for biological origin.

Ambiguities of tests for life

It is remarkable, however, that many of these tests are ambiguous. It would be possible, for example, for the Martian investigator to guess wrong about what terrestrial organisms eat and to make incorrect assumptions about their structural chemistry or their interaction medium. If forms of regular geometry that do not move were detected microscopically, there might be serious questions of biological versus mineralogical origin. Chemical criteria (such as the expectation that if odd-numbered carbon chains are more prominent than even-numbered carbon chains, then life is detected) might not be valid unless it was certain which processes actually occurred in the prebiological organic chemistry of the planet in question. In addition, there might be the galling problem of contamination. The Martians' spacecraft might carry living organisms from their own planet and report them as detected on the planet Earth. For this reason great care would have to be taken that spacecraft were rigorously sterilized.

In fact, many of these problems have already arisen in an analysis of a variety of meteorite called carbonaceous chondrites. These meteorites, which fall on the Earth probably from the asteroid belt, contain about 1 percent organic matter by mass, far too much to be largely the result of terrestrial contamination. The most abundant organic molecules, however, are not clearly of biological origin, and some of the biologically more interesting molecules may be contaminants. Reports of optical activity have been contested and might alternatively be due to contamination. Geometrically interesting microscopic inclusions have been detected in these bodies. The most abundant inclusions, however, are probably mineralogical in origin, while the most highly structured and lifelike are very rare and, at least in some cases, are obviously the result of contamination (in one case by ragweed pollen). Finally, claims have been made of the extraction of viable microorganisms from the interiors of carbonaceous chondrites. These meteorites are porous, however, and "breathe" air in and out during their entry into the atmosphere. There also have been significant opportunities for their contamination after arrival on the Earth. Moreover, one of the organisms extracted was a facultative aerobe. Since, as yet, no planet in the solar system besides the Earth is known to contain significant quantities of molecular oxygen, it seems quite curious that the complex electron-transfer apparatus required for oxygen metabolism would be evolved out on the asteroid belt in expectation of ultimate arrival on the Earth. Here, again, contamination has proved a serious hazard. The large amounts of organic matter that are found in carbonaceous chondrites, however, suggest that the production of organic molecules occurred with very great efficiency in the early history of the solar system.

From such a hypothetical exercise as the instrumental detection of life on Earth by an extraterrestrial observer and from the actual experience acquired in the analysis of carbonaceous chondrites, the following conclusions can be drawn: There is no single and unambiguous "life detector." There are instruments of great generality that make few ambiguous assumptions about the nature of extraterrestrial organisms, particularly their chemistry. These systems, however, require a fair degree of luck (an animal must walk by during the operating lifetime of the instrument), or they require the solution of difficult instrumental problems (such as the acquisition and preparation of samples for remote microscopic examination). Other instruments, such as metabolism detectors, have great sensitivity and are directed at the more abundant microorganisms. They are quite specific, however, and are critically dependent upon certain assumptions (for example, that extraterrestrial organisms eat sugars) that are no better than informed guesses. Therefore, an array of instruments, both very general and very specific, seems required. Stringent sterilization of such spacecraft appears necessary, both to avoid confusion of the life-detection experiments and to prevent interaction of contaminants with the indigenous ecology. Many of the instruments and strategies discussed in the preceding paragraphs continue to be adapted by the United States in attempts to search for life on the Moon and the nearby planets.

An exobiological survey of the solar system. A brief survey of the physical environments and biological prospects of the moons and planets of the solar system, so far as is known, follows. The Moon's surface seems inhospitable to life of any sort. The diurnal temperatures range from about 100 to about 400 K. In the absence of any significant atmosphere or magnetic field, ultraviolet light and charged particles from the Sun penetrate unimpeded to the lunar surface, delivering in less than an hour a dose lethal to the most radiation-resistant microorganism known. For other

Biological prospects of the Moon

reasons already mentioned, the absence of an atmosphere and of any liquid medium on the surface also argues against life. The subsurface environment of the Moon is not nearly so inclement. About a metre or so subsurface there is no penetration of ultraviolet light or solar protons, and the temperature is maintained at a relatively constant value about 230 K. Even there, however, the absence of an atmosphere and the probable absence of abundant liquids make the biological prospects rather dim.

It is not out of the question, however, that prebiological organic matter, produced in the early history of the Moon, might be found sequestered beneath the lunar surface. Such organic matter may have been produced either in an original lunar atmosphere that has subsequently been lost to space, or in a secondary lunar atmosphere produced by release of gases after the formation of the Moon, and also subsequently lost to space. The depth at which such organic matter may be found depends upon the unknown history of the early lunar atmosphere, if any, and upon whether the Moon has, on the whole, gained or lost matter due to meteoritic impact. An apparent gaseous emission near the lunar crater Alphonsus was recorded in 1958 and a spectral identification was made of the molecule C_2, a likely organic fragment, but this identification subsequently has been disputed.

Because of contamination by unmanned spacecraft, the lunar surface had accumulated a microbial load estimated by the late 1960s at some 100,000,000 microorganisms. Since such organisms will be immediately killed unless shielded from radiation, and since the likelihood of their growth seems remote, such contamination may not be a serious problem in subsequent microbial analysis of returned lunar samples. A much more serious contamination problem occurs during the acquisition of such samples by astronauts. Samples obtained during the historic Apollo 11 Moon landing in July 1969 were tested for possible organic molecules, but results were inconclusive. Such a finding might shed significant light on the early history of organic molecules in the solar system.

The environment of Mercury is rather like that of the Moon. Its surface temperatures range from about 100 to about 620 K, but about a metre subsurface the temperature is constant, very roughly at comfortable room temperature on Earth. But the absence of any significant atmosphere, the unlikelihood of bodies of liquid, and the intense solar radiation make life unlikely.

Speculations about life on Mars

Direct evidence for life on Mars has been claimed for many decades. The first such argument was posed by a French astronomer, E.L. Trouvelot, in 1884: "Judging from the changes that I have seen to occur from year to year in these spots, one could believe that these changing grayish areas are due to Martian vegetation undergoing seasonal changes." The seasonal changes on Mars have been reliably observed, not only visually but also photometrically. There is a conspicuous springtime increase in the contrast between the bright and dark areas of Mars. Accompanying colour changes have been reported, but their reality has been disputed. While such changes have been attributed to the growth of vegetation, seasonally variable dust storms are an equally convincing possibility.

The most famous case, historically, for life on Mars is the discovery of the "canals," a set of apparent thin straight lines that cross the Martian bright areas and extend for hundreds and sometimes thousands of kilometres. They change seasonally as do the Martian dark areas. These lines, first systematically observed by an Italian astronomer, G.V. Schiaparelli, in 1877, were further cataloged and popularized by a U.S. astronomer, Percival Lowell, around the turn of the century. Lowell argued from the unerring straightness of the lines that they could not be of geological origin but must instead be the artificial constructs of a race of intelligent Martians. He suggested that they might be channels carrying water from the melting polar caps to the parched equatorial cities of Mars. While considerable skepticism has been expressed about these straight lines, there is no doubt that approximately rectilinear features do exist on the Martian surface. More probable explanations, however, include crater chains, terrain contour boundaries, faults, mountain chains, and

Photograph, from a TIROS weather satellite, of a region near Cochrane, Ontario. The crisscross pattern in white (top left) shows logging swaths, a sign of intelligent life on Earth.

By courtesy of National Aeronautics and Space Administration

ridges analogous to the suboceanic ridge systems that are features of the Earth.

In July and August 1976, two U.S. probes bearing equipment designed to detect the presence or remains of organic material made successful landings on Mars. Analyses of atmospheric and soil samples met with procedural difficulties and yielded initially ambiguous and inconclusive results, although the data were later generally interpreted as negative, at least for the vicinity of the probe.

Features of Venus

According to both ground-based and space-borne observations, the average surface temperatures of Venus are around 750 K. It does not seem likely, either at the poles or on the tops of the highest Venus mountains, that the surface temperature will be below 400 K, and noontime temperatures are probably significantly hotter than 700 K. Thus, quite apart from the other surface conditions, the temperatures on Venus seem too hot for terrestrial life. It is still not possible to exclude a Venus surface life with a rather different chemistry, although hydrogen bonding would be much less suitable for the geometrical configuration of polymers on Venus than it is on Earth. The clouds of Venus, however, are another matter. There, carbon dioxide, sunlight, and (according to the results of the Venera space vehicles) water are to be found. These are the prerequisites for photosynthesis. Some molecular nitrogen also is expected at the cloud level, and some supply of minerals can be expected from dust convectively raised from the surface. The cloud pressures are about the same as on the surface of the Earth, and the temperatures in the lower clouds also are quite Earthlike. Despite the fact that there is little oxygen, the lower clouds of Venus are the most Earthlike extraterrestrial environment known. While there are no recorded cases of organisms on Earth that lead a completely airborne existence throughout their life cycle, it is not impossible that such organisms could exist in the vicinity of the Venus clouds, perhaps buoyed, as is a fish by its swim bladder, to avoid downdrafts carrying them to the hotter lower atmosphere.

A similar speculation can be entertained with regard to the lower clouds of Jupiter. On Jupiter the atmosphere is composed of hydrogen, helium, methane, ammonia, and probably neon and water vapour. But these are exactly those gases used in primitive-Earth simulation experiments directed toward the origin of life. Laboratory and computer experiments have been performed on the application of energy to simulated Jovian atmospheres. In addition to the immediate gas-phase products, such as hydrogen cyanide and acetylene, more complex organic molecules, including aromatic hydrocarbons, are formed in lower yield. The visible clouds of Jupiter are vividly coloured, and it is possible that their hue is attributable to such coloured organic compounds. There is also an apparent absorption feature near 2,600 Å, in the ultraviolet spectrum of Jupiter, which has been attributed both to aromatic hydrocarbons and to nucleotide bases. In any event it is likely that organic molecules are being produced in significant yield on Jupiter; it is possible that Jupiter is a vast planetary laboratory that has been operating for 5,000,000,000 years on prebiological organic chemistry.

The other Jovian planets, Saturn, Uranus, and Neptune, are similar in many respects to Jupiter, although much less is known about them. Their cloud-top temperatures progressively decrease with distance from the Sun. In the case of Saturn, microwave studies have indicated that the atmospheric temperature increases with depth below the clouds; similar situations are expected on Jupiter, Uranus, and Neptune. Thus, it is by no means clear that the low temperatures of the upper clouds of the Jovian planets apply to the lower clouds, or to the underlying atmosphere. The environment of Pluto is almost completely unknown.

Satellites and comets of the solar system

In addition to these planets, the solar system contains 32 natural satellites, some of which, such as Titan, a satellite of Saturn, and Io, a satellite of Jupiter, appear to have atmospheres. There are also tens of thousands of comets, which, judging from their spectra, contain organic molecules, as well as some thousands of asteroids and asteroidal fragments revolving about the Sun between the orbits of Mars and Jupiter. These are the presumed sources of the carbonaceous chondrites, which contain organic matter.

In short, there is a wide range of environments of biological interest within the solar system. There is no direct evidence for extraterrestrial life on these planets, but, on the other hand, there is no strong evidence against life on many of these worlds. Beyond this is the near certainty that biologically interesting organic molecules will be found throughout the solar system.

Intelligent life beyond the solar system. For thousands of years man has wondered whether he is alone in the universe or whether there might be other worlds populated by creatures more or less like himself. The common view, both in early times and through the Middle Ages, was that the Earth was the only "world" in the universe. Nevertheless, many mythologies populated the sky with divine beings, certainly a kind of extraterrestrial life. Many early philosophers held that life was not unique to the Earth. Metrodorus, an Epicurean philosopher in the 3rd and 4th centuries BC, argued that "to consider the Earth the only populated world in infinite space is as absurd as to assert that in an entire field sown with millet, only one grain will grow." Since the Renaissance there have been several fluctuations in the fashion of belief. In the late 18th century, for example, practically all informed opinion held that each of the planets was populated by more or less intelligent beings; in the early 20th century, by contrast, the prevailing informed opinion (except for the Lowellians) held that the chances for extraterrestrial intelligent life were insignificant. In fact the subject of intelligent extraterrestrial life is for many people a touchstone of their beliefs and desires, some individuals very urgently wanting there to be extraterrestrial intelligence, and others wanting equally fervently for there to be no such life. For this reason it is important to approach the subject in as unbiased a frame of mind as possible. A respectable modern scientific examination of extraterrestrial intelligence is no older than the 1950s. The probability of advanced technical civilizations in our galaxy depends on many controversial issues.

The Green Bank formula

A simple way of approaching the problem, which illuminates the parameters and uncertainties involved, has been devised by a U.S. astrophysicist, F.D. Drake. The number N of extant technical civilizations in the galaxy can be expressed by the following equation (the so-called Green Bank formula):

$$N = R_* f_p n_e f_l f_i f_c L$$

where R_* is the average rate of star formation over the lifetime of the galaxy; f_p is the fraction of stars with planetary systems; n_e is the mean number of planets per star that are ecologically suitable for the origin and evolution of life; f_l is the fraction of such planets on which life in fact arises; f_i is the fraction of such planets on which intelligent life evolves; f_c is the fraction of such planets on which a technical civilization develops; and L is the mean lifetime of a technical civilization. What follows is a brief consideration of the factors involved in choosing numerical values for each of these parameters, and an indication of some currently popular choices. In several cases these estimates are no better than informed guesses and no very great reliability should be pretended for them.

There are about 2×10^{11} stars in the galaxy. The age of the galaxy is about 10^{10} years. A value of $R_* = 10$ stars per year is probably fairly reliable. While most contemporary theories of star formation imply that the origin of planets is a usual accompaniment of the origin of stars, such theories are not well enough developed to merit much confidence. Through the painstaking measurement of slight gravitational perturbations in the proper motions of stars, it has been found that about half of the very nearest stars have dark companions with masses ranging from about the mass of Jupiter to about 30 times the mass of Jupiter. The nearest of these dark companions orbit Barnard's star, which is only six light-years from the sun and is the second nearest star system. The most direct indication that planetary formation is a general process throughout the universe is the existence of satellite systems of the major planets of our own solar system. Jupiter, with 16 satellites, Saturn with 20 or more, and Uranus with five each closely resemble miniature solar systems. It is not known what the distribution of distances of planets from their central star are in other solar systems and whether they tend to vary systematically with the luminosity of the parent star. But considering the wide range of temperatures that seem to be compatible with life, it can be tentatively concluded that $f_p n_e$ is about one.

Likelihood of origin of life

Because of the apparent rapidity of the origin of life on Earth, as implied by the fossil record, and because of the ease with which relevant organic molecules are produced in primitive-Earth simulation experiments, the likelihood of the origin of life over a period of billions of years seems high, and some scientists believe that the appropriate value of f_l is also about one. For the quantities of f_i and f_c the parameters are even more uncertain. The vagaries of the evolutionary path leading to the mammals, and the unlikelihood of such a path ever being repeated has already been mentioned. On the other hand, intelligence need not necessarily be restricted to the same evolutionary path that occurred on the Earth; intelligence clearly has great selective advantage, both for predators and for prey.

Similar arguments can be made for the adaptive value of technical civilizations. Intelligence and technical civilization, however, are clearly not the same thing. For example, dolphins appear to be very intelligent, but the lack of manipulative organs on their bodies has apparently limited their technological advance. Both intelligence and technical civilization have evolved about halfway through the relevant lifetime of the Earth and Sun. Some, but by no means all, evolutionary biologists would conclude that the product $f_i f_c$ taken as 10^{-2} is a fairly conservative estimate.

Still more uncertain is the value of the final parameter, L, the lifetime of a technical civilization. Here, fortunately for man, but unfortunate for the discussion, there is not even one example. Contemporary world events do not provide a very convincing counterargument to the contention that technical civilizations tend, through the use of weapons of mass destruction, to destroy themselves shortly after they come into being. If we define a technical civilization as one capable of interstellar radio communication, our technical civilization is only a few decades old. If then L is about 10 years, multiplication of all of the factors assumed above leads to the conclusion that there is in the second half of the 20th century only about one technical civilization in the galaxy—our own. But if technical civilizations tend to control the use of such weapons and avoid self-annihilation, then the lifetimes of technical civilizations may be very long, comparable to geological or stellar evolutionary time scales; the number of technical civilizations in the galaxy would then be immense. If it is believed that about 1 percent of developing civilizations make peace with themselves in this way, then there are about 1,000,000 technical civilizations extant in the galaxy. If they are randomly distributed in space, the distance from the Earth to the nearest such civilization will be several hundred light-years. These conclusions are, of course, very uncertain.

Communications with extraterrestrial civilizations

How is it possible to enter into communication with another technical civilization? Independent of the value of *L*, the above formulation implies that there is about one technical civilization arising every decade in the galaxy. Accordingly, it will be extraordinarily unlikely for man soon to find a technical civilization as backward as his. From the rate of technical advance that has occurred on the Earth in the past few hundred years, it seems clear that man is in no position to project what future scientific and technical advances will be made even on Earth in the next few hundred years. Very advanced civilizations will have techniques and sciences totally unknown to 20th-century man. Nevertheless man already has a technique capable of communication over large interstellar distances. This technique, already encountered in the discussion of life on Earth, is radio transmission. Imagine that we employ the largest radio telescope available on Earth, the 1,000-foot-diameter dish of Cornell University, the Arecibo Observatory in Puerto Rico, and existing receivers, and that the identical equipment is employed on some transmitting planet. How distant could the transmitting and receiving planets be for intelligible signals to be transmitted and received? The answer is a rather astonishing 1,000 light-years. Within a volume centred on the Earth, with a radius of 1,000 light-years, there are more than 10,000,000 stars.

There would of course be problems in establishing such radio communication. The choices of frequency, of target star, of time constant, and of the character of the message would all have to be selected by the transmitting planet so that the receiving planet would, without too much effort, be able to deduce the choices. But none of these problems seem insuperable. It has been suggested that there are certain natural radio frequencies (such as the 1,420-megacycle line of neutral hydrogen) that might be tuned to; the first choice might be to listen to stars of approximately solar spectral type; in the absence of a common language there nevertheless are messages whose intelligent origin and intellectual content could be made very clear without making many anthropocentric assumptions.

Because of the expectation that the Earth is relatively very backward, it does not make very much sense to transmit messages to hypothetical planets of other stars. But it may very well make sense to listen for radio transmissions from planets of other stars. Project Ozma, a very brief program of this sort, oriented to two nearby stars, Epsilon Eridani and Tau Ceti, was organized in 1960 by Drake. On the basis of the Green Bank formula, it would be very unlikely that success would greet an effort aimed at two stars only 12 light-years away, and Project Ozma was unsuccessful. It remains, however, the first pioneering attempt at interstellar communication. Related programs were organized on a larger scale and with great enthusiasm in the 1960s in the U.S.S.R., where a state scientific commission devoted to such an effort was organized. Other communication techniques including laser transmission and interstellar spaceflight have been discussed seriously and may not be infeasible, but if the measure of effectiveness is the amount of information communicated per unit cost, then radio is the method of choice.

The search for extraterrestrial intelligence is an extraordinary pursuit, in part because of the enormous significance of possible success, but in part because of the unity it brings to a wide range of disciplines: studies of the origins of stars, planets, and life; of the evolution of intelligence and of technical civilizations; and of the political problem of avoiding man's self-annihilation. But at least one point is clear. In the words of Loren Eiseley (also from *The Immense Journey*),

> Lights come and go in the night sky. Men, troubled at last by the things they build, may toss in their sleep and dream bad dreams, or lie awake while the meteors whisper greenly overhead. But nowhere in all space or on a thousand worlds will there be men to share our loneliness. There may be wisdom; there may be power; somewhere across space great instruments, handled by strange, manipulative organs, may stare vainly at our floating cloud wrack, their owners yearning as we yearn. Nevertheless, in the nature of life and in principles of evolution we have had our answer. Of men [as are known on earth] elsewhere, and beyond, there will be none forever.

BIBLIOGRAPHY. Definitions of life and life on Earth are explored in LAWRENCE J. HENDERSON, *The Fitness of the Environment: An Inquiry into the Biological Significance of the Properties of Matter* (1913, reissued 1970); PHILIP HANDLER (ed.), *Biology and the Future of Man* (1970); FRANKLIN M. HAROLD, *The Vital Force: A Study of Bioenergetics* (1986); LYNN MARGULIS and DORION SAGAN, *Microcosmos: Four Billion Years of Evolution from Our Microbial Ancestors* (1986, reprinted 1991), and *What Is Life?* (1995), valuable treatments of the microbial basis of life; RENATO DULBECCO, *The Design of Life* (1987), a study of life at the molecular level; and a special issue of *Scientific American* (Oct. 1994), devoted to life in the universe.

The origin of life is discussed in many books. FRANCES CRICK, *Life Itself: Its Origin and Nature* (1981), posits an extraterrestrial origin for life on Earth. ROBERT SHAPIRO, *Origins: A Skeptic's Guide to the Creation of Life on Earth* (1986), critiques assorted theories on the origin of life. ANDREW SCOTT, *The Creation of Life: Past, Future, Alien* (1986), provides a brief, readable account. An original cosmological argument for the origin of life is proposed in EUGENE F. MALLOVE, *The Quickening Universe: Cosmic Evolution and Human Destiny* (1987). SIDNEY W. FOX, *The Emergence of Life: Darwinian Evolution from the Inside* (1988), outlines the author's research and philosophy. A.G. CAIRNS-SMITH, *Seven Clues to the Origin of Life: A Scientific Detective Story* (1990), approaches the topic from the standpoint of thermodynamic laws and clues from the chemical operation of known cells. Also of interest are FREEMAN DYSON, *Origins of Life* (1985); GEORGE GREENSTEIN, *The Symbiotic Universe: Life and Mind in the Cosmos* (1988); and CHRISTIAN DE DUVE, *A Blueprint for a Cell: The Nature and Origin of Life* (1991), and *Vital Dust: Life as a Cosmic Imperative* (1995).

Studies on extraterrestrial life include A.G.W. CAMERON (ed.), *Interstellar Communication* (1963); I.S. SHKLOVSKII and CARL SAGAN, *Intelligent Life in the Universe* (1966; originally published in Russian, 1962); ELIE A. SHNEOUR and ERIC A. OTTESEN (compilers), *Extraterrestrial Life: An Anthology and Bibliography* (1966); COLIN S. PITTENDRIGH, WOLF VISHNIAC, and J.P.T. PEARMAN (eds.), *Biology and the Exploration of Mars* (1966); WALTER SULLIVAN, *We Are Not Alone: The Continuing Search for Extraterrestrial Intelligence*, rev. ed. (1993); SAMUEL GLASSTONE, *The Book of Mars* (1968); CARL SAGAN and JEROME AGEL, *The Cosmic Connection* (1973, reissued 1989); EDWARD REGIS, JR. (ed.), *Extraterrestrials: Science and Alien Intelligence* (1985), a collection of essays, some advocating and others dissenting from the concept of life on other planets; EMMANUEL DAVOUST, *The Cosmic Water Hole* (1991; originally published in French, 1988), highlighting European research efforts; FRANK DRAKE and DAVA SOBEL, *Is Anyone Out There? The Scientific Search for Extraterrestrial Intelligence* (1992), an account of Drake's research; PAUL DAVIES, *Are We Alone?: Philosophical Implications of the Discovery of Extraterrestrial Life* (1995), essays that both support and refute the idea; and BEN ZUCKERMAN and MICHAEL H. HART (eds.), *Extraterrestrials—Where are They?*, 2nd ed. (1995), a collection of essays, most of which are skeptical of the existence of extraterrestrial life. Histories of the theories of extraterrestrial life are found in STEVEN J. DICK, *Plurality of Worlds: The Origins of the Extraterrestrial Life Debate from Democritus to Kant* (1982); and MICHAEL J. CROWE, *The Extraterrestrial Life Debate, 1750–1900: The Idea of a Plurality of Worlds from Kant to Lowell* (1986). THOMAS R. MCDONOUGH, *The Search for Extraterrestrial Intelligence: Listening for Life in the Cosmos* (1987); and JOSEPH A. ANGELO, JR., *The Extraterrestrial Encyclopedia: Our Search for Life in Outer Space*, rev. and updated (1991), examine the technologies used in the quest for extraterrestrial life.

(C.Sn./Ed.)